

AD-A050 035

AIR WEATHER SERVICE SCOTT AFB ILL  
SELECTED TOPICS IN STATISTICAL METEOROLOGY. (U)  
JUL 77 R G MILLER, B D ALTENHOF, J N FULFORD  
AWS-TR-77-273

F/G 4/2

UNCLASSIFIED

NL

1 OF 2  
AD  
A050035



AD A050035

AD No. —  
DDC FILE COPY



AWS-TR-77-273

*16*

SELECTED TOPICS  
IN  
STATISTICAL METEOROLOGY

DDC  
RECEIVED  
FEB 15 1978  
A

Approved for public release; distribution unlimited

PUBLISHED BY  
AIR WEATHER SERVICE (MAC)  
UNITED STATES AIR FORCE  
JULY 1977



REVIEW AND APPROVAL STATEMENT

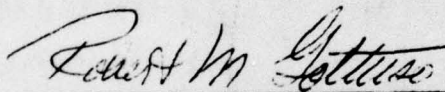
This report is approved for public release. There is no objection to unlimited distribution of this report to the public at large, or by DDC to the National Technical Information Service (NTIS).

This technical report has been reviewed and is approved for publication.



ROBERT G. MILLER  
Chief Scientist, Air Weather Service

FOR THE COMMANDER



ROBERT M. GOTTUSO, Col, USAF  
DCS/Aerospace Sciences  
Air Weather Service

Unclassified  
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

14. REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM	
1. REPORT NUMBER AWS-TR-77-273	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER	
6. TITLE (and Subtitle) SELECTED TOPICS IN STATISTICAL METEOROLOGY.		7. TYPE OF REPORT & PERIOD COVERED Final rept.	
7. AUTHOR(s) Robert G. Miller, Editor Chief Scientist, Air Weather Service		8. CONTRACT OR GRANT NUMBER(s)	
9. PERFORMING ORGANIZATION NAME AND ADDRESS Hq Air Weather Service (MAC) ✓ Scott AFB, IL 62225		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS	
11. CONTROLLING OFFICE NAME AND ADDRESS Hq Air Weather Service (MAC) Scott AFB, IL 62225		12. REPORT DATE Jul 1977	13. NUMBER OF PAGES 164
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) 12 172p.		15. SECURITY CLASS. (of this report) Unclassified	
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE			
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.			
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report) 10 Robert G. Miller.			
18. ABSTRACT (Continue on reverse side if necessary and identify by block number) <del>This report is a collection of papers which were written by graduate students at Saint Louis University, St. Louis, MO. Authors include:</del> Bruce D. Altenhof, Jimmy N. Fulford, William S. Weaving, William W. Neubert, Wendell M. Pool, Jr., Roger C. Whiton, John W. Louer, James W. Taylor;			
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Meteorology Single Station Analysis Discriminant Analysis Weather Forecasting Markov Processes Stepwise Regression Statistics Multivariate Analysis Matrices Statistical Probability Regression Analysis Delphi Technique Probability Dummy Variables Nonlinear Regression			
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) This report is a compilation of papers on statistical meteorology resulting from a graduate course at Saint Louis University on special topics in statistical meteorology. Methodologies covered include: preliminary mathematics, screening regression, multiple discriminant analysis, regression estimation of event probabilities, Markov processes, nonlinear prediction, Delphi technique, and the results of a single station forecasting experiment. Numerous examples are included. Appendixes cover computer programs on → next page			



Unclassified

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

Continued from Block #18:


Robert G. Miller; Michael J. Kelly, Jr., and Jeanette M. Heumann.

Continued from Block #19:

Canonical correlation      Nonparametric Statistics      Simulation

Continued from Block #20:

*cont* → screening predictors, Crout matrix operations, and REEP.



### PREFACE

The material contained in this report is a compilation of students' papers coming out of my one semester graduate course at Saint Louis University. Each of the papers is an expansion of a class lecture or topic. For various reasons, not all of the papers submitted to satisfy course requirements have been included.

The objective of the course was to develop insight into ways of applying statistics to solve problems. It was not intended to be a course on statistics. As a result, the papers are not a comprehensive coverage of these subjects. Furthermore, they are not the last word on the subject. They are not authoritative coverages of a subject, but more a description or interpretation of the subject by the student. However, they have been reviewed for technical content. The intention is to develop interest on the part of the reader to do further investigating into statistics.

There was some effort made to keep the papers primarily on the subject of applied meteorology. However, since the course included applications of statistics to other fields, there are two papers which cover other subjects. These papers cover applications which discuss techniques that can and, in some cases, have been directly applied to meteorology elsewhere.

For those wishing some guidelines for more basic probability and statistical material, I would recommend, in addition to the texts referred to in the bibliography, books by: Wadsworth and Bryan, Hogg and Craig, Dixon and Massey. An appendix includes copies of computer programs for doing some of the analyses discussed in the report.

Time restrictions have made it impractical to properly edit the enclosed papers for uniformity of content. The report has the character of a compilation of preprints to a proceeding with each paper reflecting its own style of presentation.

ACCESSION for	
NTIS	White Section <input checked="" type="checkbox"/>
ODC	Buff Section <input type="checkbox"/>
UNANNOUNCED	<input type="checkbox"/>
JUSIFICATION	
BY	
DISTRIBUTION/AVAILABILITY CODES	
Dist.	SPECIAL
A	



### ACKNOWLEDGEMENTS

Several computing facilities were used during the class experiment on single-station forecasting. It is a pleasure to thank the following people and organizations:

#### MILITARY AIRLIFT COMMAND (MAC)

Col Castor Mendez-Vigo, Jr  
Col Kurt G. Blunck  
Lt Howard D. Campbell  
Mr Leonard A. Mehmert

#### DEFENSE COMMERCIAL COMMUNICATIONS OFFICE (DECCO)

Col Stanley P. Houghton  
Capt Larry Hawkins  
Mr Oliver Banks

#### ENVIRONMENTAL TECHNICAL APPLICATIONS CENTER (ETAC)

Col Paul Janota  
Lt Col James A. Degiovanni  
Capt Donald S. Thomas (from Carswell AFB)  
TSgt Marvin Freimund

#### SAINT LOUIS UNIVERSITY INFORMATION SYSTEMS

Mr Fred Ravetta

This document would not have appeared so quickly without the typing assistance of Dolores Bates, Mary Jean Biehn, Mildred Collins, Jane Holtgrave, Wilma Kell, Irene Maus, Mary Ann Myers, Mildred Revoire, and Helen Severit, all at Headquarters, Air Weather Service.

Editing help by Col Robert M. Gottuso, Capt Gary K. Dotson, and Mr Walter Burgmann is gratefully appreciated.

## CONTENTS OF REPORT

PREFACE	v
ACKNOWLEDGEMENTS	vi
CONTENTS OF REPORT	vii

### CHAPTER

1	INTRODUCTION	Maj Bruce D. Altenhof
2	PRELIMINARY MATHEMATICS	Capt Jimmy N. Fulford
3	SCREENING REGRESSION	Capt William S. Weaving
4	MULTIPLE DISCRIMINANT ANALYSIS	Capt William W. Neubert
5	REGRESSION ESTIMATION OF EVENT PROBABILITIES	Capt Wendell M. Pool, Jr
6	CANONICAL CORRELATION APPLICATIONS	Lt Jeanette M. Heumann
7	MARKOV PROCESSES	Capt Roger C. Whiton
8	NONLINEAR PREDICTION	Mr John W. Louer
9	DELPHI TECHNIQUE	Lt Col James W. Taylor
10	RESULTS OF A SINGLE STATION FORECASTING EXPERIMENT	Dr Robert G. Miller Capt Roger C. Whiton Capt Michael J. Kelly, Jr

### BIBLIOGRAPHY

### APPENDICES

## CHAPTER 1

### INTRODUCTION

by Maj Bruce D. Altenhof

#### 1.1 Introduction.

This publication is the result of a St. Louis University graduate-level course on special topics in Statistical Meteorology taught by Adjunct Professor, Dr. Robert G. Miller, in the 1977 spring semester. Dr. Miller, a meteorological statistician, is the Chief Scientist for the Air Weather Service. The objective of this text, like that of the course, is to create statistical insight into how to solve problems.

Two textbooks were used to supplement Dr. Miller's lecture material. The books were "Multivariate Analysis: Techniques for Educational and Psychological Research" written by Maurice M. Tatsuoka (1971) and "Statistics: A Guide to the Unknown" edited by Judith M. Tanur and by Frederick Mosteller, William H. Kruskal, Richard F. Link, Richard S. Pieters, Gerald R. Rising (1972). Tatsuoka's book is based on courses he taught in advanced-statistics and multivariate-analysis. It provides the additional information required by students to get a thorough understanding of advanced statistical methods. Tanur's book explores ways in which statistics can be applied to a variety of problem areas in society. The book contains 44 nontechnical examples of applied statistics and the contributions made in all aspects of today's society (government, business, science, etc).

In an effort to stimulate student interest and involvement, a class experiment was undertaken on single-station forecasting and is included as a chapter.

#### 1.2 Statistical Subjects.

Numerous statistical subjects were explained during the course. The main subjects in alphabetical order were:

- (1) Analysis of Variance
- (2) Analysis of Covariance
- (3) Bayes' Theorem
- (4) Canonical Correlation
- (5) Clustering
- (6) Decision Theory
- (7) Markov Processes
- (8) Monte Carlo
- (9) Multiple Discriminant Analysis
- (10) Multivariate Normal Distributions
- (11) Nonparametric Procedures
- (12) Orthogonal Polynomials
- (13) Pattern Recognition
- (14) Regression Analysis
- (15) Significance Testing
- (16) Simulation Techniques
- (17) Stochastic Processes
- (18) Transformations

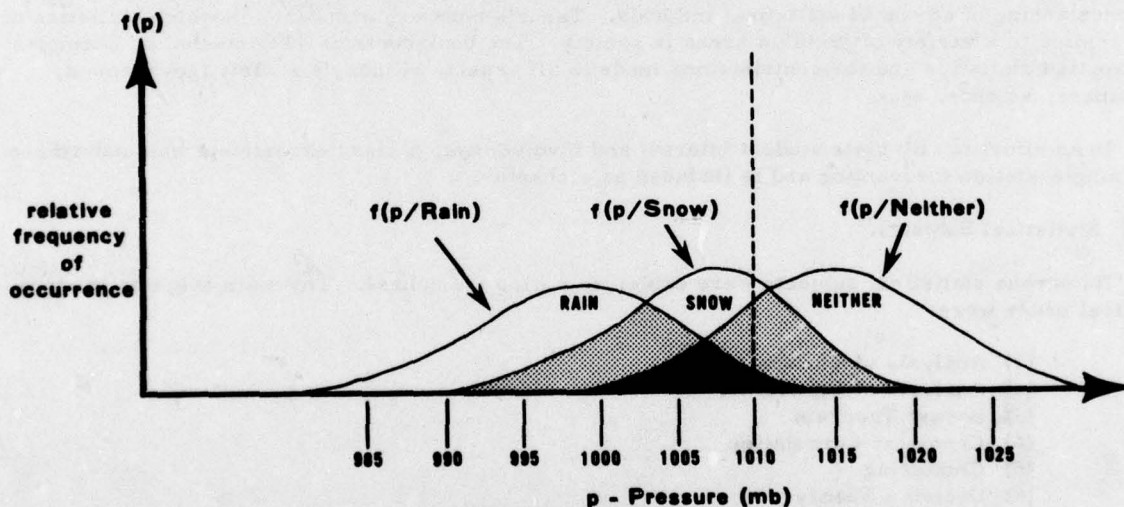
The Analysis of Variance (ANOVA) technique is a method for testing hypotheses concerning means of several populations. The ANOVA technique was illustrated by depicting its application to testing the relative merits of a new YMCA running program as compared to the program it replaced. Two test groups were analyzed, one used the old YMCA program and the other used the new running program. The analysis tested the mean values of the dependent variable (time to run a certain distance) for the two groups. In doing so, ANOVA allowed one to conclude whether there was a difference between the two programs.



The statistical technique known as the analysis of covariance is an extension of the ANOVA to take into account the possible effects, on the dependent variable, of one or more uncontrolled variables (the covariates). In the above test of the YMCA running programs, it was determined that the new program appeared to be better than the older program because the mean of the variable considered (time to run a certain distance) was much lower under the new program. After the test was conducted, it was determined that the age of the people tested was not uniform. Age in this case was a covariate and may have had an effect on the dependent variable. The analysis of covariance was then applied to permit an adjustment to be made--to sharpen the test results. The analysis of covariance uses regression equations to make estimates of the dependent variable from the known values of the predictor variable, which in this case was age.

Bayes' theorem on inverse probability or *posterior* probability was depicted by the following examples: As a weatherman, suppose you have to make a forecast of the precipitation conditions (rain, snow, or neither) at the inauguration in Washington DC one day in advance. All you have is your trusty barometer. For years you have recorded the pressure on each Jan 19th, 9 AM EST, and later took notice of what occurred the next day at the inauguration site. If you plotted the relative frequency of occurrence of the precipitation conditions with respect to pressure, you might get what is depicted in Figure 1-1.

FIGURE 1-1



From climatology, you know that the probability of rain ( $P_R$ ) is .20, probability of snow ( $P_S$ ) is .30, and the probability of neither occurring ( $P_N$ ) is .50. Bayes' Theorem can be written for this situation as:

$$P(\text{Rain/Press}) = \frac{P_R \cdot f(\text{press/Rain})}{P_R \cdot f(\text{press/Rain}) + P_S \cdot f(\text{press/Snow}) + P_N \cdot f(\text{press/Neither})}$$

$$P(\text{Snow/Press}) = \frac{P_S \cdot f(\text{press/snow})}{P_R \cdot f(\text{press/Rain}) + P_S \cdot f(\text{press/Snow}) + P_N \cdot f(\text{press/Neither})}$$

$$P(\text{Neither/Press}) = \frac{P_N \cdot f(\text{press/Neither})}{P_R \cdot f(\text{press/Rain}) + P_S \cdot f(\text{press/Snow}) + P_N \cdot f(\text{press/Neither})}$$



Now to determine the probabilities. You measure the pressure on 19 Jan at 9 AM EST and it is 1010mb. From the graph you determine that the f's for 1010 mb are as follows:

$$f(1010/R) = .20$$

$$f(1010/S) = .50$$

$$f(1010/N) = .40$$

Then:

$$P(\text{Rain}/1010\text{mb}) = \frac{(.20)(.20)}{(.20)(.20) + (.30)(.50) + (.50)(.40)} = \frac{.04}{.39} \approx .10$$

$$P(\text{Snow}/1010\text{mb}) = \frac{(.30)(.50)}{.39} = \frac{.15}{.39} \approx .40$$

$$P(\text{Neither}/1010\text{mb}) = \frac{(.50)(.40)}{.39} = \frac{.20}{.39} \approx .50$$

Your forecast based on the use of Bayes' Theorem would be 10% chance of rain, 40% chance of snow, and 50% of no precipitation for the inauguration.

In another example, suppose someone hands you a red ball and asks for the probabilities that the ball was taken from urn 1, urn 2, or urn 3. You are given the following information:

URN 1	URN 2	URN 3
Contains	Contains	Contains
50 red balls	90 red balls	25 red balls
50 black balls	10 black balls	75 black balls

You also know *a priori* that the probability it came from Urn 1 is 1/3, from Urn 2 is 1/3, and from Urn 3 is 1/3.

By applying Bayes' Theorem and using the above information, the probabilities are easily determined by

$$P(\text{Urn } x | \text{red}) = \frac{P_{\text{urn } x} \cdot P(\text{red} | \text{Urn } x)}{\sum_{x=1}^3 P_{\text{urn } x} \cdot P(\text{red} | \text{Urn } x)} \quad x = 1, 2, 3$$

Therefore,

$$P(\text{Urn 1/red}) = \frac{(1/3)(.50)}{(1/3)(.50) + (1/3)(.90) + (1/3)(.25)} = \frac{.50}{1.65} \approx .30$$

$$P(\text{Urn 2/red}) = \frac{(1/3)(.90)}{(1/3)(.50) + (1/3)(.90) + (1/3)(.25)} = \frac{.90}{1.65} \approx .54$$

$$P(\text{Urn 3/red}) = \frac{(1/3)(.25)}{(1/3)(.50) + (1/3)(.90) + (1/3)(.25)} = \frac{.25}{1.65} \approx .16$$

The statistical method of Canonical Correlation was created by an economist in 1935. It is a method by which one determines a linear combination of  $p$  predictors and a linear combination of  $q$  criterion variables such that the correlation between these linear combinations in the total sample is as large as possible. This technique has been used successfully in the insurance business. To illustrate, let's consider the following example--Buyer vs Product. An insurance company will look at all of its policies sold (Product) and lists them under categories such as policy size, policy type, mode of payment, etc. The company then lists all the factors known about the buyers of the product (e.g., age, sex, income, employment, etc). From a complete canonical correlation analysis of all variables, the company can determine, from the factors known about a potential customer, which type of policy, amount of insurance, and payment plan the customer is most likely to purchase. Its use in determining what policies a customer is most likely to buy saves them considerable time, money and manpower. Another possible area of use for this method is in meteorology. One should be able to infer tomorrow's set of weather variables from today's set of weather variables.

Clustering is a method for identifying significant groups such as would be desired in market research; e.g., What are the main markets that a company is operating in successfully? Again, insurance companies have found this information extremely valuable for sales training, designing new products, etc. In meteorology, clustering can be used in weather typing.

Decision Theory is a technique that is applied when there is uncertainty. The following example will provide insight into this technique. Suppose you were planning the next day's work for a construction team. You have a choice of three projects you can plan for; however, it depends on whether it rains or doesn't rain as to what project you can accomplish. From past experience, you know what losses you would incur if the weather doesn't agree with the project you planned for. By putting this information in a decision matrix and applying decision theory, you can minimize the potential for loss. Consider the following matrix.

Units of Loss		States of Nature	
		RAIN	NO RAIN
Planning Action	A <sub>1</sub>	10	0
	A <sub>2</sub>	4	4
	A <sub>3</sub>	0	10

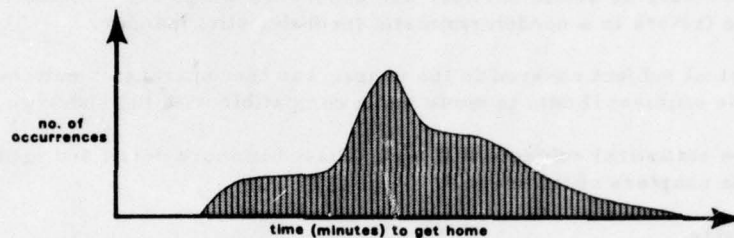
If the forecast for the next day was 100% for rain, you would plan for project 3 or A<sub>3</sub>. If the forecast was for 100% no rain, then you would plan for project 1 or A<sub>1</sub>. However, there is usually some uncertainty as to whether it will rain or not. For instance, there may be a 50% chance of rain. By using the probability of rain in conjunction with the matrix, one arrives at the following:

	RAIN	NO RAIN	
Probability	.5	.5	
A <sub>1</sub>	10	0	<p>EXPECTED LOSSES</p> <p>= .5 · 10 + .5 · 0 = 5 units lost</p> <p>= .5 · 4 + .5 · 4 = 4 units lost</p> <p>= .5 · 0 + .5 · 10 = 5 units lost</p>
A <sub>2</sub>	4	4	
A <sub>3</sub>	0	10	

Based on this decision theory technique, you would plan for Project A<sub>2</sub> because it minimizes your expected loss. Since at different times other probabilities for rain may be forecast (i. e., 10%, 20%, 65%, 90%, etc), each can be entered into your calculations to minimize the potential loss to your organization.

The "Markov Process" is a statistical approach where knowing the last state of a process tells you all you need to know to predict a future state.

A technique often used in statistics is the "Monte Carlo" method. The Monte Carlo technique is applied to stochastic processes which are based on a large number of variable factors. An example of a process which is affected by many variables is the time it takes an individual to drive home after work. A few of the variables are: time of day, traffic condition, weather, and the individual's attitude--each of which has a wide range of variability. By keeping records over a long period of time, one can graph the number of occurrences of particular times it takes to get home.



Using the technique of Monte Carlo, you can develop a model for simulating the time distribution in place of collecting the information from actual experience.

Multiple Discriminant Analysis, or MDA, as it is frequently referred to, involves a more complicated application (more variables) of Bayes' Theorem. In this method, the variables are weighted so that the groups separate from each other thus allowing the probabilities to become "sharper."

The bivariate normal distribution is an extension of the univariate normal distribution to the bivariate situation. For a univariate normal distribution of a variable X, every point on a line represents a possible value of X. For the normal bivariate distribution of variable X and variable Y, every point on a plane represents a possible pair of values. Application of the bivariate normal distribution has been made in hurricane predictions.

Most statistical methods require assumptions about the distributions underlying a model. Non-parametric methods, also called "distribution-free" statistical methods, are used for making inferences without any assumption as to the form of distribution in the population.

Statistical methods using polynomials have been developed to reduce large data arrays to ones of manageable size. It is possible, because of redundancies in data, to reduce the number of observations needed to represent a given situation. Harmonic functions and empirical orthogonal functions are used to accomplish this data reduction also. Orthogonal polynomials have been used to represent weather maps with only a few numbers.

Classification problems have long interested statisticians. "Pattern Recognition" is one approach for determining which of several groups a particular individual "resembles" the most, in terms of a specific set of measurable characteristics. In pattern recognition, one has at hand a sample from each of K well-defined populations. Associated with each individual are measurements on P variables that are deemed to be important in differentiating among the several populations or groups. Now, we take a new individual whose group membership is unknown, but for whom we can take measures on the same P variables. Using the measures, one can classify him as a member of one of these K groups according to which he shows greatest resemblance.

"Regression" is the estimation, or prediction, of an unknown value of one variable from known values of one or more other variables. In the simplest case, one variable is predicted from a known value of another variable. The known variable is commonly called the independent



variable and the unknown variable, the one to be estimated, is called the dependent variable. The relationship between the dependent variable and independent variable is given by regression equations.

Significance testing is another important area of statistical analysis. It arises in a special way through the use of screening procedures which try to find the best predictor variables.

Simulations have wide applications in statistics and various techniques will be discussed in the text. Simulations are accomplished through modeling. By analyzing past data, models can be developed which are used to simulate actual observed conditions. Once the model has been developed, one can input into the model various parameters and determine or simulate what the future (or predicted) effects will be of the input parameter on the model. The Monte Carlo method is a simulation technique.

Stochastic processes, as stated earlier, are processes which are affected by a very large number of variable factors in a nondeterministic (probabilistic) manner.

Another statistical subject covered in the course was that of transformations. The analysis of data can be made efficient if data is made more compatible with the underlying models.

All of the above statistical subjects will be discussed in more detail and applied in various combinations in the chapters of this report.

### 1.3 Chapter Contents.

This report contains 10 chapters. Each represents the individual student's efforts to expand and add details to the statistical methods presented by Dr. Miller during the class sessions. Several chapters deal with a particular operational problem solved by employing statistical methods.

Chapter 2 provides the preliminary mathematics for the study of statistical methods. It gives an explanation on the presentation of data bases used for most statistical analysis work. Matrices are illustrated since data are normally presented in this format. Problems with data handling, such as missing or erroneous data, are discussed. Two methods of handling these problems are shown. They are the prime number scheme and the error vectors with the logical "OR" statement. The concept of random variables is applied to statistical editing. The purpose and history of transformations are highlighted, followed by a discussion on the transformation to dummy variables. Detailed instructions are provided for the Crout reduction procedure, the derivation of the auxiliary matrix from both symmetrical and nonsymmetrical matrices, and the calculation of the inverse matrix. With a working knowledge of these preliminaries, one should be able to understand the basic principles which make up the various methods of statistical analysis discussed in the report.

In chapter 3, the topic of screening regression is covered. The screening or stepwise procedure is a method of selecting significant independent variables and determining a rank order listing of significant independent variables as related to a dependent variable. This technique has application in meteorology where the independent variables are often considered as predictors and the dependent variable is the predictand. The screening process allows one to find which of a large number of possible predictors are most significant and uses these predictors in the regression equation to predict future meteorological conditions. Background information on the origin and development of the screening process is highlighted. The details of the screening procedure, the tests of significance, and several applications are thoroughly covered.

Chapter 4 discusses the application of multiple discriminant analysis (MDA) to precipitation forecasting. The rationale for the use of MDA versus regression is covered. The chapter provides insight into graphical interpretation, mathematical procedures, selection of predictors, and estimation of probabilities.



Chapter 5 looks at the use of Regression Estimation of Event Probabilities (REEP) as a forecasting technique. The REEP prediction technique is discussed and an experiment using the technique is examined.

Market Research is covered in Chapter 6 with an in-depth discussion of canonical correlation and its use in determining the relationship between the buyer and the product. This type of statistical information is a valuable management tool.

In chapter 7, a very detailed and thorough analysis of Markov processes and its use in meteorology are presented. The chapter provides a review of matrix manipulations and concepts. The Markov chain is defined. A comparison of the equivalent Markov model developed by Dr. Miller is compared to the classical Markov model and is found to be much easier to develop and apply to practical forecasting problems. The chapter shows that the simple yet powerful prediction methods of the Markov type can successfully be applied to the problem of forecasting the weather at a station, given only an initial observation from that station. This is known as single-station forecasting.

Chapter 8 considers the problem of nonlinearity and a method for dealing with it. Boolean algebra and a property of lattice theory are used to uncover nonlinear relationship.

The Delphi technique is explained in chapter 9. The Delphi technique is defined and its method of application is presented. A detailed discussion of the Delphi process, as it was applied to a class exercise, is provided as an excellent example of the technique.

In chapter 10, there is a write-up on the class experiment. The experiment applied a statistical method to develop probability weather forecasts for Rickenbacker AFB. The verification on independent data is presented.

## CHAPTER 2

### PRELIMINARY MATHEMATICS

by CAPT JIMMY N. FULFORD

#### DATA NOTATION

Most statistical analyses work from data bases. This data is normally presented in a matrix format with the variables along the top and the observations along the side; e.g.:

#### A TYPICAL DATA ARRAY

	<u>VARIABLES</u>						
Observation Number	$x_1$	$x_2$	$x_3$	.	.	.	$x_p$
1	84	5	10	.	.	.	24
2	17	7	13	.	.	.	4
3	32	4	16	.	.	.	35
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
N	37	3	12	.	.	.	19

= M

Underlining signifies that M is a matrix. The variables,  $x_i$  ( $i = 1, \dots, p$ ), can be used for items such as temperature, pressure, visibility, etc., and their values could be the hourly observations of these variables at a weather station in raw or coded form.

#### PROBLEMS WITH DATA

When building a matrix of data there are certain problems associated with collecting and storing it. Raw data seldom comes in a neat form. There can be missing observations, questionable or illogical information as well as other gross errors. Passing through the data by eye or with a computer one can find missing or gross errors if the amount of data is small enough. Corrections could then be made in most instances. If there is a large volume of data, the following two methods could be used to keep track of missing data. One is the prime number scheme and the other uses error vectors with the logical "OR" statement. Suppose that the following matrix represented  $p$  variables such as temperature, dewpoint, pressure, etc., with  $N$  hourly observations (zeros signifying missing data).

	<u>VARIABLES</u>						
Observation Number	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	.	$x_p$
1	27	17	4	5	17	.	97
2	34	0	0	13	8	.	34
3	0	21	2	7	0	.	86
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
N	29	17	13	4	15	.	42

Assign a prime number to each variable such as

$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	.	.	.	.	$x_p$
2	3	5	7	11	13	.	.	.	.	$N_p$

2-1

Create a Prime Vector whose elements are equal to the product of the prime numbers associated with missing data in a given row of observations.

PRIME VECTOR		$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_p$
$a_1 = 0$	1	27	17	4	5	17	97
$a_2 = 15$	2	34	0	0	13	8	34
$a_3 = 22$	3	0	21	2	7	0	86
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
$a_i = 13$	N	29	17	13	4	15	42

In the example shown, the first observation row has no missing or grossly erroneous data; therefore, element  $a_1$  of the prime vector is equal to zero. The second element,  $a_2$ , of the prime vector is equal to 15. This is the product of prime members 3 and 5; therefore,  $x_{22}$  and  $x_{23}$  of the data matrix are missing or erroneous. Element  $a_3$  of the prime vector is equal to 22 or the product of prime numbers 2 and 11. This means that  $x_{31}$  and  $x_{35}$  of the data matrix are missing or erroneous. This computational method to keep track of missing data is easy to carry along as an extra vector for gross editing for later reference for building a "good" sample of required variables.

Another method uses the same size matrix as the data matrix but assigns zeros and ones for good and bad data. For such a situation the results or "bits" can be packed on the computer word. Zeros are placed in the elements when the data is good and ones are used when the data is missing or incorrect. Then a resultant vector is generated using a logical "OR" -- Example:

	$e_1$	$e_2$
1	0	0
2	0	0
3	0	1
.	.	.
.	.	.
.	.	.
N	0	0

$e_1$  and  $e_2$  are the error vectors associated with variables  $x_1$  and  $x_2$ .

The symbol for logical "OR" is  $\oplus$ . The logic is as follows:

	$e_1$	
	$\oplus$	
	0	1
	0	
$e_2$	1	
	1	1

If either  $e_1$  or  $e_2$  or both are equal to one, then the output is one. The output is zero only if both  $e_1$  and  $e_2$  are zero. In the example shown, observations 1 and 2 of  $x_1$  and  $x_2$  are good but observation 3 cannot be used if variable  $x_2$  is required.

These two methods are examples of gross and expedient methods. If a thorough edit is desired, then it should be performed statistically. In order to accomplish a statistical edit, it is necessary to be able to treat each variable numerically; that is, all qualitative variables must be converted to random variables. To do this, each observed condition must be assigned a number. NOTE: A Random Variable is a variable having a specified range of values with definite probabilities associated with each. An example would be present weather which consists of qualitative variables,



e.g., fog, snow, rain, thunderstorm, different cloud types, etc. For ease of operation, each of these transformations can be given the values 0 and 1--referred to as dummy variables. Quantitative variables may also need to be transformed; i.e., changing wind from degrees and speed to "u" and "v" components.

## TRANSFORMATIONS

The purpose and history of transformations will be discussed before discussing the transformation of dummy variables.

### Purpose of Transformations

Analysis of data will proceed easier if the effects are additive, and the variability of error is symmetrical and near normal. The purpose of transformations is to approach these properties as nearly as possible. Normally a transformation which improves one of the properties also improves the other.

### History

If  $N$  items are each drawn independently and at random from an infinite population where a proportion  $P$  of the items have a property  $A$ , then the proportions  $P_1, P_2, P_3, \dots$  of items possessing  $A$  in the successive samples will be distributed in such a manner that, as the number of independent samples increases without limit, the average of the  $P$ 's will approach  $P$  and the mean of their squared deviations from  $P$  will approach  $P(1-P)/N$ . In the language of statistics this is expressed by stating that, if a sample of  $N$  items is drawn at random from an infinite population in which a proportion  $P$  of the items have an attribute  $A$ , and if the proportion of items possessing  $A$  in the sample is denoted by  $p$ , then:

$$\text{Expected value of } P = E(p) = P \quad (2-1)$$

$$\text{Variance of } P = V(p) = \frac{P(1-P)}{N} \quad (2-2)$$

$$\text{Standard Deviation of } P = \sigma(p) = \sqrt{\frac{P(1-P)}{N}} \quad (2-3)$$

Equation (2-3) or equation (2-2), considered along with equation (2-1), states that observed proportions  $P$ , based on successive independent random samples of size  $N$ , may be expected to be grouped more closely about the true proportion  $P$  when  $N$  is large than when  $N$  is small. For fixed sample size, their sampling variation about  $P$  will be greatest when  $P$  equals  $1/2$  and will decrease toward zero as either  $P$  or  $1-P$  approaches zero.

### Dummy Variables

As stated earlier, in order to accomplish a statistical edit, qualitative variables must be replaced as numbers. Dummy variables are normally designated by the letter  $Z$ . Table 2-1 shows how the meteorological variable ceiling is transformed into five dummy variables which categorizes ceilings into the height intervals shown.

TABLE 2-1

OBS NO.	CEILING Feet	DUMMY VARIABLES				
		$Z_1$	$Z_2$	$Z_3$	$Z_4$	$Z_5$
		$\leq 100$	200-400	500-900	1000-2900	$\geq 3000$
1	Unlimited	0	0	0	0	1
2	10,000	0	0	0	0	1
3	5,000	0	0	0	0	1
4	2,000	0	0	0	1	0
5	7,000	0	0	0	0	1
6	0	1	0	0	0	0
7	400	0	1	0	0	0
.	.					
.	.					
.	.					
N	Unlimited	0	0	0	0	0

As indicated in Table 2-1, ceiling is grouped into five classes. There are no values for numbers such as 99 or 499 in these limits because ceiling is measured to the nearest hundred feet. Whenever



the ceiling is  $\leq 100$  feet, as in observation 6, then dummy variable 1 is assigned a value of one and dummy variables 2, 3, 4, and 5 are set equal to 0. Whenever the ceiling is in the range 200-400 feet, as in observation 7, then dummy variable 2 takes on a value of 1 and dummy variables 1, 3, 4, and 5 are equal to 0. If there are N observations of ceiling, then there will be N observations of each of the five dummy variables.

If a continuous variable such as pressure or temperature is to be expressed with dummy variables, then it is necessary to separate the continuous values into classes. D. R. Cox(1957) devised a method for dividing a continuous variable into K classes so that the grouping error is minimized under certain conditions for a stated K. This method computes K-1 limits and then assigns dummy variables exactly as before. The limits are computed as follows: Let  $X_i$ ,  $i = 1, 2, \dots, N$ , be the N values of the variable to be transformed to K dummy variables. Calculate:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$$

$$\sigma = \left[ \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 \right]^{1/2}$$

K-1 limits are then obtained by using these values to enter Table 2-2. For example, if three dummy variables are desired (K=3) and  $\bar{X} = 1$ ,  $\sigma = 2$  then the two limits are  $1 - .612 \times 2$  and  $1 + .612 \times 2$  or  $-.224$  and  $+2.224$ . Obviously, there is some loss of resolution but the pay-off is a tremendous increase in speed of computation if all variables are dummy variables. For a variable such as temperature, it would perhaps take six divisions (K=5) to include the necessary resolution. This would transform one integer  $X_1$  into six dummy variables  $Z_i$ ,  $i = 1, \dots, 6$ .

TABLE 2-2  
FACTORS FOR DETERMINING LIMITS

K	Limits								
	$L_1$	$L_2$	$L_3$	$L_4$	$L_5$	$L_6$	$L_7$	$L_8$	$L_9$
2	$\bar{X}$								
3	$\bar{X} - 0.612\sigma$	$\bar{X} + 0.612\sigma$							
4	$\bar{X} - 0.980\sigma$	$\bar{X}$	$\bar{X} + 0.980\sigma$						
5	$\bar{X} - 1.230\sigma$	$\bar{X} - 0.395\sigma$	$\bar{X} + 0.395\sigma$	$\bar{X} + 1.230\sigma$					
6	$\bar{X} - 1.449\sigma$	$\bar{X} - 0.660\sigma$	$\bar{X}$	$\bar{X} + 0.660\sigma$	$\bar{X} + 1.449\sigma$				
7	$\bar{X} - 1.611\sigma$	$\bar{X} - 0.875\sigma$	$\bar{X} - 0.280\sigma$	$\bar{X} + 0.280\sigma$	$\bar{X} + 0.875\sigma$	$\bar{X} + 1.611\sigma$			
8	$\bar{X} - 1.748\sigma$	$\bar{X} - 1.050\sigma$	$\bar{X} - 0.500\sigma$	$\bar{X}$	$\bar{X} + 0.500\sigma$	$\bar{X} + 1.050\sigma$	$\bar{X} + 1.748\sigma$		
9	$\bar{X} - 1.887\sigma$	$\bar{X} - 1.225\sigma$	$\bar{X} - 0.715\sigma$	$\bar{X} - 0.232\sigma$	$\bar{X} + 0.232\sigma$	$\bar{X} + 0.715\sigma$	$\bar{X} + 1.225\sigma$	$\bar{X} + 1.887\sigma$	
10	$\bar{X} - 1.980\sigma$	$\bar{X} - 1.334\sigma$	$\bar{X} - 0.840\sigma$	$\bar{X} - 0.407\sigma$	$\bar{X}$	$\bar{X} + 0.407\sigma$	$\bar{X} + 0.840\sigma$	$\bar{X} + 1.334\sigma$	$\bar{X} + 1.980\sigma$

## STATISTICAL ANALYSES

To do a thorough editing or to do most statistical analysis, the following quantities are required:

### SUMS

$$X_{1,1} + X_{1,2} + \dots + X_{1,N} = \sum_{i=1}^N X_{1,i}$$

$$X_{2,1} + X_{2,2} + \dots + X_{2,N} = \sum_{i=1}^N X_{2,i}$$

$$\dots$$

$$X_{p,1} + X_{p,2} + \dots + X_{p,N} = \sum_{i=1}^N X_{p,i}$$

### SUM OF SQUARES

$$X_{1,1}^2 + X_{1,2}^2 + \dots + X_{1,N}^2 = \sum_{i=1}^N X_{1,i}^2$$

over all p variables:

### SUM OF CROSS PRODUCTS

$$X_{1,1}X_{2,1} + X_{1,2}X_{2,2} + \dots + X_{1,N}X_{2,N} = \sum_{i=1}^N X_{1,i}X_{2,i}$$

over all pairs leading to the elements of the following matrix:

$$\underline{X} = \begin{bmatrix} \sum_{i=1}^N X_{0,i}^2 & \sum_{i=1}^N X_{0,i}X_{1,i} & \sum_{i=1}^N X_{0,i}X_{2,i} & \dots & \sum_{i=1}^N X_{0,i}X_{p,i} \\ \sum_{i=1}^N X_{0,i}X_{1,i} & \sum_{i=1}^N X_{1,i}^2 & \sum_{i=1}^N X_{1,i}X_{2,i} & \dots & \sum_{i=1}^N X_{1,i}X_{p,i} \\ \sum_{i=1}^N X_{0,i}X_{2,i} & \sum_{i=1}^N X_{1,i}X_{2,i} & \sum_{i=1}^N X_{2,i}^2 & \dots & \sum_{i=1}^N X_{2,i}X_{p,i} \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{i=1}^N X_{0,i}X_{p,i} & \sum_{i=1}^N X_{1,i}X_{p,i} & \sum_{i=1}^N X_{2,i}X_{p,i} & \dots & \sum_{i=1}^N X_{p,i}^2 \end{bmatrix}$$

Where  $X_{0,i} = 1, i = 1, \dots, N$ . Thus  $\sum_{i=1}^N X_{0,i}^2 = N$ .

The above matrix is sufficient to perform most statistical analyses. A reasonable sample where  $N = 10,000$  observations and  $P = 100$  variables would require over 50,000,000 multiplications (taking account of symmetry in  $\underline{X}$ ). There are ways to avoid this large amount of computation. One is by the use of dummy variables previously shown and the other is by the use of screening methods which will be discussed in another section. Now we will examine a method of editing one of the  $X_i$  variables where  $i = 1, P$ . The particular  $X_i$  will be expressed as  $Y$ . It is possible to estimate the value of  $Y$  from the other  $P-1$  variables such as

$$\hat{Y} = B_0X_0 + B_1X_1 + \dots + B_{p-1}X_{p-1} \quad (2-4)$$

where  $Y$  is omitted from the  $X_i$ ,  $i = 1, \dots, P$  variables, such that  $\sum_{i=1}^N (Y_i - \hat{Y}_i)^2$  is a minimum.  $\hat{Y}$  is defined as  $Y$  estimate and  $X_0 \equiv 1$ . Notice that with the inclusion of  $X_0$ , the above equation now has  $P$  terms since the particular  $X_i = Y$  is not included. In order to determine equation (2-4), it will be advantageous to employ the Crout method.

#### CROUT METHOD

The Crout method is a modification of the Gauss reduction. It is well suited to the use of desk calculators and electronic computers. In addition, the storage of auxiliary data is reduced.

The following data are used to derive the initial matrix  $I$ .  $Z$ 's instead of  $X$ 's are used since the variables are assumed to be in dummy form.

Variables  $Z_i$ ,  $i=1, \dots, 3$ .

Observation Number	VARIABLES			
	$Z_0$	$Z_1$	$Z_2$	$Z_3 = Y$
1	1	1	0	0
2	1	1	1	1
3	1	1	0	1
4	1	0	1	0
5	1	1	1	1
6	1	0	0	0
7	1	0	0	0
8	1	0	1	1
9	1	0	0	0
10	1	1	0	0

Thus,

$$\sum_{i=1}^{10} Z_{0,i}^2 = 10$$

$$\sum_{i=1}^{10} Z_{1,i} Z_{2,i} = 2$$

$$\sum_{i=1}^{10} Z_{2,i} Y_i = 3$$

$$\sum_{i=1}^{10} Z_{0,i} Z_{1,i} = \sum_{i=1}^{10} Z_{1,i}^2 = 5$$

$$\sum_{i=1}^{10} Z_{0,i} Y_i = \sum_{i=1}^{10} Y_i^2 = 4$$

$$\sum_{i=1}^{10} Z_{0,i} Z_{2,i} = \sum_{i=1}^{10} Z_{2,i}^2 = 4$$

$$\sum_{i=1}^{10} Z_{1,i} Y_i = 3$$

$$O = \begin{bmatrix} 10 & & & \\ & 5 & 5 & \\ & 4 & 2 & 4 \\ & 4 & 3 & 3 & 4 \end{bmatrix}$$

If a set of linear equations were being solved, then  $O$  would represent the initial matrix in the Crout method. This will simplify the computations since the cross-product matrix is symmetrical, and is a special case of the general Crout solution. A second example of the complete Crout method will be shown after the symmetrical case.

#### Auxiliary Matrix

The first step is to derive from a given initial matrix  $O$ , an auxiliary matrix  $A$ . As the steps in deriving the auxiliary matrix progress, each step is dependent upon the preceding steps. The auxiliary matrix is evolved in a right-angle pattern branching from the main diagonal, i.e., the



diagonal starts at the upper left-hand corner and slopes downward to the right. To distinguish between the usual terms, column and row, which denote the entire vertical and horizontal arrays, we shall use terms for partial columns and rows. Therefore, we define a vertical block (Symbol:  $V_t$ ,  $t = 1, 2, \dots, n$ ) as that part of a column and the elements between it. A horizontal block (Symbol:  $H_t$ ,  $t = 1, 2, \dots, n$ ) is defined as that part of a row which lies to the right of the diagonal element. These definitions are illustrated schematically in the following diagram.

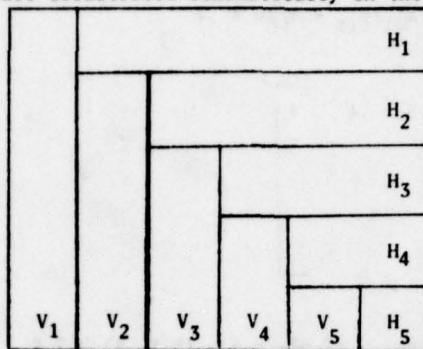


DIAGRAM 2-1

The auxiliary matrix is completed in successive stages beginning with the first vertical block  $V_1$  and then forming the first horizontal block  $H_1$ , next  $V_2$  and then  $H_2$ , next  $V_3$  and then  $H_3$ , and continuing in a similar fashion throughout, alternating from vertical to horizontal and then to the next vertical and horizontal --  $V_1, H_1, V_2, H_2, \dots, V_n, H_n$ . These blocks are formed according to the following rules:

$$a'_{ij} = a_{ij} - \sum_{k=1}^{j-1} a'_{ik} a'_{kj} \quad (i \geq j) \quad (2-5)$$

$$a'_{ij} = \frac{a'_{ji}}{a'_{ii}} \quad (i < j) \quad (2-6)$$

$$\underline{0} = \begin{bmatrix} a_{11} & & & \\ a_{21} & a_{22} & & \\ a_{31} & a_{32} & a_{33} & \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} 10 & & & \\ 5 & 5 & & \\ 4 & 2 & 4 & \\ 4 & 3 & 3 & 4 \end{bmatrix}$$

$$\underline{A} = \begin{bmatrix} a'_{11} & a'_{12} & a'_{13} & a'_{14} \\ a'_{21} & a'_{22} & a'_{23} & a'_{24} \\ a'_{31} & a'_{32} & a'_{33} & a'_{34} \\ a'_{41} & a'_{42} & a'_{43} & a'_{44} \end{bmatrix}$$

$\underline{A}$  is the auxiliary matrix to be determined by equations (2-5) and (2-6) using the previously stated order.

In column  $V_1$ ,  $a'_{ij} = a_{ij}$  since  $j-1 = 0$ . Using equation 2-5

$$\underline{A} = \begin{bmatrix} 10 & & & \\ 5 & & & \\ 4 & & & \\ 4 & & & \end{bmatrix}$$

Row  $H_1$  uses equation (2-6)

$$a'_{12} = a_{21}/a'_{11} = 5/10 = .5$$

$$a'_{13} = a_{31}/a'_{11} = 4/10 = .4$$

$$a'_{14} = a_{41}/a'_{11} = 4/10 = .4$$

$$\underline{A} = \begin{bmatrix} 10 & .5 & .4 & .4 \\ 5 & & & \\ 4 & & & \\ 4 & & & \end{bmatrix}$$

Using equation (2-5) - Column  $V_2$

$$a'_{22} = a_{22} - a'_{21}a'_{12} = 5 - (5 \times .5) = 2.5$$

$$a'_{32} = a_{32} - a'_{31}a'_{12} = 2 - (4 \times .5) = 0$$

$$a'_{42} = a_{42} - a'_{41}a'_{12} = 3 - (4 \times .5) = 1$$

$$\underline{A} = \begin{bmatrix} 10 & .5 & .4 & .4 \\ 5 & 2.5 & & \\ 4 & 0 & & \\ 4 & 1 & & \end{bmatrix}$$

Using equation (2-6) - Row  $H_2$

$$a'_{23} = a_{32}/a'_{22} = 0/2.5 = 0$$

$$a'_{24} = a_{42}/a'_{22} = 1/2.5 = .4$$

$$\underline{A} = \begin{bmatrix} 10 & .5 & .4 & .4 \\ 5 & 2.5 & 0.0 & .4 \\ 4 & 0 & & \\ 4 & 1 & & \end{bmatrix}$$

Using equation (2-5) - Column  $V_3$

$$a'_{33} = a_{33} - a'_{31}a'_{13} - a'_{32}a'_{23} = 4 - (4 \times .4) - (0 \times 0) = 2.4$$

$$a'_{43} = a_{43} - a'_{41}a'_{13} - a'_{42}a'_{23} = 3 - (4 \times .4) - (1 \times 0) = 1.4$$

$$\underline{A} = \begin{bmatrix} 10 & .5 & .4 & .4 \\ 5 & 2.5 & 0.0 & .4 \\ 4 & 0.0 & 2.4 & \\ 4 & 1 & 1.4 & \end{bmatrix}$$

Using equation (2-6) - Row  $H_3$

$$a'_{34} = a'_{43}/a'_{33} = 1.4/2.4 = .58333$$

$$\underline{A} = \begin{array}{cccc} 10 & .5 & .4 & .4 \\ 5 & \underline{2.5} & 0.0 & .4 \\ 4 & 0.0 & \underline{2.4} & .58333 \\ 4 & 1 & 1.4 & \end{array}$$

Using equation (2-5) - Column  $V_4$

$$a'_{44} = a_{44} - a'_{41}a'_{14} - a'_{42}a'_{24} - a'_{43}a'_{34}$$

$$a'_{44} = 4 - (4 \times .4) - (1 \times .4) - (1.4 \times .58333) = 1.183$$

$$\begin{array}{cccc} 10 & .5 & .4 & .4 \\ 5 & \underline{2.5} & 0.0 & .4 \\ 4 & 0.0 & \underline{2.4} & .58333 \\ 4 & 1 & 1.4 & \underline{1.183} \end{array}$$

It can be shown that  $a'_{44} = 1.183 = \sum_{i=1}^N (Y_i - \hat{Y}_i)^2$ .

The solution to the coefficients of the estimate equation  $\hat{Y} = B_0 + B_1Z_1 + B_2Z_2$  is as follows:

$$B_2 = a'_{34} = .58333$$

$$B_1 = a'_{24} - a'_{23}B_2 = .4 - (0 \times .58333) = .4$$

$$B_0 = a'_{14} - a'_{12}B_1 - a'_{13}B_2$$

$$B_0 = .4 - (.5 \times .4) - (.4 \times .58333) = -.0333$$

The above procedure for calculating coefficients will also apply for continuous variables.

To edit the dummy variable observations examine  $|Y - \hat{Y}|$ . If it is, say,  $>.99$  reject a value of zero for  $Y$ . If it is, say,  $<.01$  reject a value of one for  $Y$ .

Observation	VARIABLES			Y	$\hat{Y}$	$[Y - \hat{Y}]$
	$Z_0$	$Z_1$	$Z_2$			
1	1	1	0	0	.367	.367
2	1	1	1	1	.950	.050
3	1	1	0	1	.367	.633
4	1	0	1	0	.550	.550
5	1	1	1	1	.950	.050
6	1	0	0	0	-.033	.033
7	1	0	0	0	-.033	.033
8	1	0	1	1	.550	.450
9	1	0	0	0	-.033	.033
10	1	1	0	0	.367	.367



Another feature of the auxiliary matrix that can be used for a statistical analysis, if the variables are continuous, is with

$$\sigma^2 = \sum_{i=1}^N (Y_i - \hat{Y}_i)^2 / N.$$

#### Determinant

The determinant of the initial matrix is equal to the product of the diagonal elements of the auxiliary matrix. The determinant of our initial matrix is therefore equal to  $10 \times 2.5 \times 2.4 \times 1.183 = 70.98$ .

Note: Since  $\hat{Y}$  is the probability that  $Y = 1$  you may use the following procedure to eliminate values of  $\hat{Y} > 1$  or  $< 0$ .

If  $\hat{Y} < 0$  set  $\hat{Y} = 0$ .

If  $\hat{Y} > 1$  set  $\hat{Y} = 1$ .

#### Non-Symmetrical Matrix

The procedure to determine the auxiliary matrix for a non-symmetrical matrix and the steps for finding the inverse will now be shown. Suppose we have the following linear equations:

$$2x + 3y + z = 8$$

$$x + 4y + z = 7$$

$$x + y + 2z = 5$$

The coefficient matrix is as follows:

$$\underline{O} = \begin{bmatrix} 2 & 3 & 1 \\ 1 & 4 & 1 \\ 1 & 1 & 2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

The augmented matrix is as follows:

$$\underline{M} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{matrix} c_1 \\ c_2 \\ c_3 \end{matrix} \quad \underline{O} | \underline{C} = \begin{bmatrix} 2 & 3 & 1 & 8 \\ 1 & 4 & 1 & 7 \\ 1 & 1 & 2 & 5 \end{bmatrix}$$

Using the coefficient matrix as our initial matrix, we then use the following equations to determine the auxiliary matrix:

$$a'_{ij} = a_{ij} - \sum_{k=1}^{j-1} a'_{ik} a'_{kj} \quad (i \geq j) \quad (2-7)$$

$$a'_{ij} = \frac{1}{a_{ii}} \left[ a_{ij} - \sum_{k=1}^{i-1} a'_{ik} a'_{kj} \right] \quad (i < j) \quad (2-8)$$

Equation (2-7) is identical to equation (2-5) used previously. Equation (2-8) is different since the initial matrix is non-symmetrical. The format and order of determining the elements of  $\underline{A}$  are identical to the previous case. Therefore, a detailed explanation is not given but by using equations (2-7) and (2-8)  $\underline{A}$  is as follows:

$$\underline{A} = \begin{bmatrix} 2 & 1.5 & .5 \\ 1 & 2.5 & .2 \\ 1 & -.5 & 1.6 \end{bmatrix}$$

The next step is to determine the C column which will then be used to determine the solution column  $\underline{X}$  whose elements are the required values of x, y, z or using different notation  $x_1, x_2, x_3$ . The following two equations are used to determine the solution:

$$c'_i = \frac{1}{a_{ii}} \left[ c_i - \sum_{K=1}^{i-1} a'_{iK} c'_K \right] \quad (2-9)$$

$$x_i = c'_i - \sum_{K=i+1}^N a'_{iK} x_K \quad (2-10)$$

Using equation (2-9)

$$c'_1 = c_1 / a_{11} = 8 / 2 = 4$$

$$c'_2 = (c_2 - a_{21}c'_1) / a_{22}$$

$$c'_2 = (7 - (1 \times 4)) / 2.5 = 1.2$$

$$c'_3 = (c_3 - a_{31}c'_1 - a_{32}c'_2) / a_{33}$$

$$c'_3 = (5 - (1 \times 4) - (-.5 \times 1.2)) / 1.6 = 1$$

Using equation (2-10)

$$x_3 = c'_3 = 1$$

$$x_2 = c'_2 - a_{23}x_3 = 1.2 - (.2 \times 1) = 1$$

$$x_1 = c'_1 - a_{12}x_2 - a_{13}x_3 = 4 - (1.5 \times 1) - (.5 \times 1) = 2$$

Therefore the solution to the original set of equations is  $\underline{X} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}$

#### Inverse Matrix Calculation

The procedure for determining the inverse matrix is to set  $c_k = 1$  and all the other c values = 0 where  $k = 1, \dots, 3$ . Then the  $\underline{x}_k$  column becomes the kth column of the inverse matrix. In the previous example the  $\underline{c}$  columns now become

$$\underline{C} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Using the first column of  $\underline{C}$   $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$  and equations (2-9) and (2-10):

$$c'_1 = 1/2 = .5$$

$$c'_2 = (0 - (1 \times .5)) / 2.5 = -.2$$

$$c'_3 = (0 - (1 \times .5) - (-.5 \times -.2)) / 2.5 = -.375$$

$$x_3 = -.375$$

$$x_2 = -.2 - (.2 \times -.375) = -.125$$

$$x_1 = .5 - (1.5 \times -.125) - (.5 \times -.375) = .875$$

Using the second column  $\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$  and equations (2-9) and (2-10) yields the following:

$$c_1' = 0 \quad x_3 = .125$$

$$c_2' = .4 \quad x_2 = .375$$

$$c_3' = .125 \quad x_1 = -.625$$

Using the third column  $\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$  and equations (2-9) and (2-10) yields the following:

$$c_1' = 0 \quad x_3 = .625$$

$$c_2' = 0 \quad x_2 = -.125$$

$$c_3' = .625 \quad x_1 = -.125$$

The inverse matrix is therefore:

$$(.125) \begin{bmatrix} 7 & -5 & -1 \\ -1 & 3 & -1 \\ -3 & 1 & 5 \end{bmatrix} = A^{-1}$$

and,

$$x_1 = .875c_1 - .625c_2 - .125c_3$$

$$x_2 = -.125c_1 + .375c_2 - .125c_3$$

$$x_3 = -.375c_1 + .125c_2 + .625c_3$$

giving  $x_1 = 2$ ,  $x_2 = 1$ , and  $x_3 = 1$ .

The term .125 or 1/8 was common to all elements in the inverse and was therefore factored out. The determination of an inverse matrix is particularly desirable when the set is to be solved for many distinct sets of right-hand members. The inverse matrix is also useful in getting the standard error of the regression coefficients (see Snedecor, 1946).



## CHAPTER 3

### SCREENING REGRESSION

by CAPT WILLIAM S. WEAVING

#### CONTENTS

1. Introduction
2. Linear Regression
3. Multiple Regression
4. Screening Regression
5. Applications of Screening Regression
6. Conclusion

#### 1. Introduction

The screening (or stepwise) procedure is a means of selecting significant independent variables and determining a rank order listing of these variables as related to a dependent variable. In meteorology these independent variables ( $X_m$ ,  $m = 1, \dots, M$ ) are often considered as predictors and the dependent variable ( $Y$ ) as the predictand. Thus, the stepwise procedure will consider, in turn, all the possible predictors by running tests on each independent variable and selecting the most significant variables. The procedure orders the independent variables by selecting the most significant first, the next most significant next, and so forth. The order is not necessarily optimum. The dependent variables are then used in the form of a multiple regression equation to predict the expected value of a particular dependent variable ( $Y$ ).

The number of variables considered to forecast phenomena such as severe storms, pressure patterns, hurricanes and other meteorological variables is often considerable. The desired approach is to find which of the possible predictors is most significant and use them in the regression equation to predict future conditions of ( $Y$ ). The screening process allows us to reach this goal.

Other reasons for using the screening process are as follows:

- 1) Meteorologists usually have a "feel" for the predictors which are most significant, and screening regression can be used to confirm this selection of variables based on the data.
- 2) Equations with fewer variables are easier to understand and hence more likely to gain acceptance and to be used.
- 3) A subset of variables can provide a better prediction equation than the full set, even though the full set has a higher multiple correlation coefficient ( $R$ ). The primary reason for this is that after you've considered a number of variables, any increase in the number of variables used may only increase the amount of shrinkage on independent data. Thus, a point is reached where the shrinkage occurs faster than  $R$  increases. It is therefore best to quit screening before this point is reached.

Prior to initiating the screening process, the set of possible independent variables should be shown to relate to the dependent variable(s). Experience or preliminary investigations can determine this. This is necessary because the use of regression analysis to "find" relationships, where no physical facts show the relationship to exist, frequently leads to less than desirable results.

Also remember that the final equation is a result of the data base used to develop it. Changes in the data base (e.g., additional data becomes available, new predictors are found, etc) will require periodic recomputation of the regression equation(s). Similarly, different equations will

often arise when considering different time steps (i.e., 12, 24 or 48 hours in the future), indicating that some predictors are stronger than others for these various time steps.

The screening technique is not a new idea; it has its origins back in the early 1940's with J. G. Bryan. It was further developed in the mid and late 1950's by M. A. Efroymsen and R. G. Miller. More recently, in the mid to late 1960's and early 1970's, it has been used in meteorology by the Techniques Development Laboratory (TDL) with M. A. Alaka, et al, the Air Force Global Weather Central (AFGWC) with R. C. Miller and others, the National Severe Storms Forecast Center (NSSFC) with C. L. David and others. Many other groups have applied the procedure to a great many meteorological elements. Also, applications have not just been confined to the field of meteorology; many other fields have also used screening regression. However, these other applications are beyond the scope of this paper. I am sure you can visualize the utilization of the screening technique in such fields as insurance, the stock markets, and public opinion polling, to mention only a few.

Before we can consider the screening procedure itself, we must first consider some of the basic concepts such as simple linear regression and multiple linear regression. The next few pages are devoted to these preliminaries. Following that, we consider the screening procedure itself, test of significance, and applications.

## 2. Linear Regression

First, consider the simple linear regression equation consisting of only one independent variable (X) and one dependent variable (Y), where,

$$\hat{Y} = a + bX \quad (3-1)$$

Here the hat (^) represents estimated values.

This equation represents the best fitting straight line for a set of points regressing Y on X. Of course, a is the intercept of the Y axis and b is slope of the line described by the equation. By describing this line as the line of best fit for a set of data points, we have either visually (for simple cases) or mathematically attempted to minimize the deviations of the points from the line. Thus the resultant linear equation gives us the best prediction of Y for a given value of X. Actually, the criterion of goodness or best fit that is employed is the principle of least squares. Here the best fitting line is that one which minimizes the sum of squares of the deviations of the observed values of Y from those predicted. Expressed mathematically, we wish to minimize

$$SSE = \sum_{i=1}^n (\hat{Y}_i - Y_i)^2 \quad (3-2)$$

where SSE has the common name Sum of Squares for Error, and  $\hat{Y}_i$  is the estimated value determined from the linear equation and  $Y_i$  is an observed value for observation i.

For this simple linear equation, the estimated value of b (the slope) for a sampling of data points can be found by solving the following equation:

$$b = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{n \sum_{i=1}^n X_i Y_i - \left( \sum_{i=1}^n X_i \right) \left( \sum_{i=1}^n Y_i \right)}{n \sum_{i=1}^n X_i^2 - \left( \sum_{i=1}^n X_i \right)^2} \quad (3-3)$$

Then  $a = \bar{Y} - b\bar{X}$  where the bar ( $\bar{X}$  and  $\bar{Y}$ ) refers to mean values of observed data. Knowing the estimated values of a and b, we can now determine estimated values of Y for given values of X.

## 3. Multiple Regression

The multiple regression equation quickly becomes complex when a large number M of independent variables ( $X_m$ ,  $m = 1, \dots, M$ ) are considered. Computations considering just two variables (one independent and one dependent) can easily be done on a desk calculator if the data sampling isn't too large. However, when considering the many variables which may be used in multiple regression problems, the number of computations soon becomes overwhelming. The modern digital computer and the "canned" multiple regression computer programs have eliminated the severe restriction on the number of variables which can be used.

A typical multiple regression equation would have the following form:

$$\hat{Y} = a + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_mx_m + \dots + b_Mx_M \quad (3-4)$$

Here M would be the number of predictors considered. It could also have a form similar to:

$$\hat{Y} = a + b_1x_1 + b_2x_2 + b_3x_1^2 + b_4x_2^2 + b_5x_1x_2 \quad (3-5)$$

This is still considered a linear model, since the term linear means that the model is linear with respect to the coefficients ( $b_j$ ). The regression coefficients  $b_j$  ( $j = 1, 2, 3, \dots, m, \dots, M$ ) are determined using the method of least squares.

When determining coefficients for a multiple regression equation, the method of least squares can be applied in several ways. For example, the Crout technique could be used with dummy variables as described in other papers in this report (see Chapters 2 and 7). The Gaussian elimination technique could also be used.

#### 4. Screening Regression

Textbooks such as Draper and Smith (1966) discuss several screening procedures, for example: (a) all possible regressions, (b) backward elimination, (c) forward selection, (d) stepwise regression and others. However, descriptions and comparisons of the various methods will be left to the textbooks. Since screening is an improved version of the forward-selection procedure, we will consider it in more detail. Miller (1962) defines this procedure as:

"The method of selecting predictors in a forward stepwise manner from a set of possible predictors where the criterion for selection is the partial correlation coefficient."

The screening procedure begins by selecting the individual independent variable which is the "best" predictor, namely as the predictor that maximizes the correlation coefficient. The correlation coefficient squared is the proportion of the variation explained by the predictor. Next, the screening procedure adds the variables to the equation sequentially, in order of importance.

At each step, the variable added is the one which increases the explained sum of squares (and, hence,  $R^2$ ) or equivalently reduces the residual sum of squares by the largest amount. The "best" set of variables may not be included in the final equation as a result of this procedure. However, the procedure provides an efficient method of developing "good" regression equations.

Remember, however, that when you select and test single predictors, some of these single predictors will be unselected, while together they might contain significant information. Therefore, you may want to make significance tests on combined variables as well.

However, Miller (1962) states that,

"In practice, . . . it has been found that  $F^*$  ( $F$  when predictors are selected) tends to work well as a significance test only when variables are considered singly. This may be a consequence of the fact that a bias is introduced in the regression coefficients as a result of selection."

We are now ready to select the first predictor which we will label  $X^{(1)}$  as described by Miller (1958) and Miller (1962). First, compute the total sum of squares of deviations from the mean (SST) for variates  $Y$  and  $X_m$  ( $m = 1, \dots, M$ ) and the total sum of products of deviations from the means (SPT) for variates  $Y$  and  $X_m$  ( $m = 1, \dots, M$ ). Mathematically the above is described as:

$$SST(Y) = \sum_{i=1}^n (Y_i - \bar{Y})^2, \quad (3-7)$$

$$SST(X_m) = \sum_{i=1}^n (X_{mi} - \bar{X}_m)^2, \quad m = 1, \dots, M \quad (3-8)$$

$$SPT(YX_m) = \sum_{i=1}^n (Y_i - \bar{Y})(X_{mi} - \bar{X}_m) \quad m = 1, \dots, M \quad (3-9)$$



Note that (3-7) and (3-8) are related to the variances of  $Y$  and  $X_m$  respectively; i.e.,  $\text{Var}(Y) = \frac{\text{SST}(Y)}{n}$  and  $\text{Var}(X_m) = \frac{\text{SST}(X_m)}{n}$ . Also (3-9) is similar to the covariance of  $Y$  and  $X_m$ , i.e.,  $\text{Cov}(YX_m) = \frac{\text{SPT}(YX_m)}{n}$ . Here, as before,  $n$  represents the sample size considered and  $m$  represents the predictor being considered. In Miller (1958) the multiple correlation coefficient ( $R^2$ ) is described as:

$$R_{YX_m}^2 = \frac{\text{Cov}(YX_m)^2}{\text{Var}(Y)\text{Var}(X_m)} \quad (3-10)$$

This was the criterion for selection used in that paper. However, in Miller (1962) the equivalent criterion for selection is described.

$$R_{YX_m}^2 = \frac{\text{Cov}(YX_m)^2}{\text{Var}(Y)\text{Var}(X_m)} = \frac{\text{SSF}(X_m)}{\text{SSR}(X_m)} \quad (3-11)$$

Where  $\text{SSF}(X_m)$  is the fitted sum of squares for  $Y$  on predictor  $X_m$  and  $\text{SSR}(X_m)$  is the residual sum of squares after fitting  $Y$  with predictor  $X_m$ . The values for these two are determined by the following formulas:

$$\text{SSF}(X_m) = \frac{\text{SPT}(YX_m)^2}{\text{SST}(X_m)} \quad (3-12)$$

and

$$\text{SSR}(X_m) = \text{SST}(Y) - \text{SSF}(X_m). \quad (3-13)$$

Now use the following criterion to select the first predictor from all possible ( $M$ ) predictors:

$$\frac{\text{SSF}(X^{(1)})}{\text{SSR}(X^{(1)})} > \frac{\text{SSF}(X_m)}{\text{SSR}(X_m)} \quad \text{for all } m = 1, \dots, M \quad (3-14)$$

Since this criterion is a function of the test statistic, the usual  $F$  ratio cannot be used to test the significance of  $X^{(1)}$ . Usually, it would be the 95% level  $F$  expressed as,

$$F_{.95} = F_{(1 - \frac{1}{20})} \quad (3-15)$$

But in the stepwise or screening procedure, the 95% level is

$$F^*_{.95} = F_{(1 - \frac{1}{20P})}, \quad (3-16)$$

where  $P$  is the number of predictors  $M$ .

Remember that each time a variable is selected as a significant predictor, the value of  $P$  decreases by one and, thus, the probability level of  $F^*_{.95}$  should be adjusted accordingly. This is significant when a small number of predictors are considered or as  $P$  approaches one.

Therefore, the predictor  $X^{(1)}$  is considered significant if

$$(n-2) \frac{\text{SSF}(X^{(1)})}{\text{SSR}(X^{(1)})} > F^*_{.95}(1, n-2), \quad (3-17)$$

where the numbers 1 and  $n-2$  within the parentheses represent the degrees of freedom. If, by this test,  $X^{(1)}$  is deemed significant, it then becomes the first of  $r$  predictors to be selected from the original set of  $M$  possible predictors. Then the screening technique searches for the next significant predictor. If, however,  $X^{(1)}$  is not significant, no predictors are selected.

To select the second significant predictor  $X^{(2)}$ , the following calculations are required:

$$SPT(X^{(1)}X_m) = \sum_{i=1}^n (X_i^{(1)} - \bar{X}^{(1)})(X_{mi} - \bar{X}_m), \quad (3-18)$$

here  $m = 1, \dots, M$  but  $X_m \neq X^{(1)}$ .

The criterion for choosing  $X^{(2)}$  is

$$\frac{SSF(X^{(2)}|X^{(1)})}{SSR(X^{(2)}|X^{(1)})} > \frac{SSF(X_m|X^{(1)})}{SSR(X_m|X^{(1)})} \quad \text{for all } m \quad X_m \neq X^{(1)} \quad (3-19)$$

Remember that the line between the variables means that you are finding SSF of  $X^{(2)}$  given  $X^{(1)}$ . Here,

$$SSF(X_m|X^{(1)}) = \begin{bmatrix} SPT(YX^{(1)}) \\ SPT(YX_m) \end{bmatrix} \cdot \begin{bmatrix} SST(X^{(1)}) & SPT(X^{(1)}X_m) \\ SPT(X^{(1)}X_m) & SST(X_m) \end{bmatrix}^{-1} \cdot \begin{bmatrix} SPT(YX^{(1)}) \\ SPT(YX_m) \end{bmatrix} - SSF(X^{(1)}), \quad (3-20)$$

and

$$SSR(X_m|X^{(1)}) = SST(Y) - SSF(X^{(1)}) - SSF(X_m|X^{(1)}) \quad (3-21)$$

Values for  $SSF(X^{(2)}|X^{(1)})$  and  $SSR(X^{(2)}|X^{(1)})$  are found in a similar fashion.

This second predictor is judged significant if

$$(N-3) \frac{SSF(X^{(2)}|X^{(1)})}{SSR(X^{(2)}|X^{(1)})} > F^*_{.95}(1, n-3) \quad (3-22)$$

Remember that the mathematical symbol  $[ ]'$  represents the transpose of the matrix or vector within brackets and  $[ ]^{-1}$  represents the inverse of the matrix within brackets.

We can now set up the pattern for the selection of all other significant predictors  $X^{(s)}$  in the general form:

$$\frac{SSF(X^{(s)}|X^{(1)} \dots X^{(s-1)})}{SSR(X^{(s)}|X^{(1)} \dots X^{(s-1)})} > \frac{SSF(X_m|X^{(1)} \dots X^{(s-1)})}{SSR(X_m|X^{(1)} \dots X^{(s-1)})} \quad (3-23)$$

where  $m=1, \dots, M$  but  $X_m \neq X^{(1)} \dots X^{(s-1)}$ ,

$$SSF(X_m|X^{(1)} \dots X^{(s-1)}) = \begin{bmatrix} SPT(YX^{(1)}) \\ \vdots \\ SPT(YX^{(s-1)}) \\ SPT(YX_m) \end{bmatrix} \cdot \begin{bmatrix} SST(X^{(1)}) & \dots & SPT(X^{(1)}X^{(s-1)}) & SPT(X^{(1)}X_m) \\ \vdots & \ddots & \vdots & \vdots \\ SPT(X^{(s-1)}X^{(1)}) & \dots & SST(X^{(s-1)}) & SPT(X^{(s-1)}X_m) \\ SPT(X^{(s-1)}X^{(1)}) & \dots & SPT(X^{(s-1)}X_m) & SST(X_m) \end{bmatrix}^{-1} \cdot \begin{bmatrix} SPT(YX^{(1)}) \\ \vdots \\ SPT(YX^{(s-1)}) \\ SPT(YX_m) \end{bmatrix} - SSF(X^{(1)} \dots X^{(s-1)}), \quad (3-24)$$

$$\begin{bmatrix} \text{SST}(X^{(1)}) & \dots & \text{SPT}(X^{(1)}X^{(s-1)}) & \text{SPT}(X^{(1)}X_m) \\ \vdots & & \vdots & \vdots \\ \text{SPT}(X^{(1)}X^{(s-1)}) & \dots & \text{SST}(X^{(s-1)}) & \text{SPT}(X^{(s-1)}X_m) \\ \text{SPT}(X^{(1)}X_m) & \dots & \text{SPT}(X^{(s-1)}X_m) & \text{SST}(X_m) \end{bmatrix}^{-1} \cdot \begin{bmatrix} \text{SPT}(YX^{(1)}) \\ \vdots \\ \text{SPT}(YX^{(s-1)}) \\ \text{SPT}(YX_m) \end{bmatrix} - \text{SSF}(X^{(1)}) - \dots - \text{SSF}(X^{(s-1)}|X^{(1)} \dots X^{(s-2)}),$$

and

$$\text{SSR}(X_m|X^{(1)} \dots X^{(s-1)}) = \text{SST}(Y) - \text{SSF}(X^{(1)}) - \text{SSF}(X^{(2)}|X^{(1)}) - \dots - \text{SSF}(X_m|X^{(1)} \dots X^{(s-1)}) \quad (3-25)$$

The Predictor  $X^{(s)}$  is significant if:

$$\frac{[N - (S + 1)] \text{SSF}(X^{(s)}|X^{(1)} \dots X^{(s-1)})}{\text{SSR}(X^{(s)}|X^{(1)} \dots X^{(s-1)})} > F_{.95}[1, N - (S + 1)] \quad (3-26)$$

Finally, the point is reached where the rest of the predictors or combinations thereof don't show any significance and the screening procedure is terminated.

##### 5. Applications of Screening Regression

As mentioned earlier, many different fields have made extensive use of the screening procedure. However, since we are primarily interested in meteorological applications, we will mention only a few of these meteorological applications in this section.

The purpose of this section is not to fully describe a few experiments but to tell you what the experiment considered and direct you to the appropriate material if you desire to study these experiments in more detail.

Three of the earlier experiments are described in studies in Statistical Weather Prediction (1958) with Thomas F. Malone as project director. R. G. Miller was the author of the first experiment. Here, the desired effect was to determine the predictability of several weather elements, 24 hours in advance at a number of stations. Altogether, seven weather variables were considered at each of 48 stations in the U. S. Thus, each of the variables at all 48 stations were considered in developing prediction equations for one or more predictands for each of the many stations tested.

The second experiment was authored by K. W. Veigas, R. G. Miller, and G. M. Howe. This experiment considered the "Probabilistic Prediction of Hurricane Movements by Synoptic Climatology." Selected hurricanes and tropical storms that occurred between 1928 and 1953 were used as the developmental sample from which the prediction equations were derived. A total of 447 storms were considered. Ninety-five variables were considered for each of the storms. Ninety-one of these variables were grid pressure values. Two of the four remaining variables were position coordinates for time ( $t=0$ ) and the other two were position coordinates 24 hours prior to prediction time. The screening regression procedure then determined which of the 95 variables were the strongest predictors. This experiment concluded that "the surface pressure pattern does contain a useful amount of information about the future movement of tropical cyclones for the subsequent twenty-four hour period."

The third experiment was authored by R. G. Miller and G. M. Howe and considered the "Statistical Prediction of the 500-mb Pattern" over North America during January and February 1957. One 24-hour prediction equation was derived for each of the 46 points used in the predictand grid by the screening regression technique. This experiment concluded that, "The predictions, which from an operational point of view are easily derived, produced results with errors of the same general magnitude as those of the JNWP barotropic model in use at the time of comparison."



R. C. Miller (1972), while working at AFGWC, used the screening procedure in a study of 328 tornado cases and found 14 significant parameters for forecasting severe storms. These parameters are contained in his Technical Report 200 and listed in rank order.

David (1973) used screening regression to develop two period equations providing an early morning determination of the severe thunderstorm potential for the day. He used Model Output Statistics (MOS) and screening regression to study the occurrence of severe thunderstorms within a radius of 120 nm of 32 stations through the central and eastern United States. For predictors he used the PE model forecast data transmitted via teletype as FOUS bulletins and he used the observed 0600Z surface data. For the first period equation, 12-, 18-, and 24-hour forecasts from the PE model for each of the following predictors were screened: mean relative humidity for 3 layers, 6 hourly quantitative precipitation totals, vertical velocity at 700 mb, the lifted index, 1000-500 mb thickness, u and v components of the mean wind of the boundary layer, mean potential temperature and mean pressure of the boundary layer. For the second period, the only difference was that the 24-, 30-, and 36-hour forecasts from the PE model were used. David concluded his report by simply stating that "... predictors from the PE model are very useful in forecasting areas of expected severe thunderstorm."

TDL, under M. A. Alaka, et al (1973), also described linear regression equations they developed in their initial attempt to forecast the likelihood of thunderstorms or severe weather in the central, eastern, and southern United States. Initially, they ran three experiments to develop medium-range prediction equations (6-24 hours). All three experiments used 24-hour forecasts from the National Meteorological Center (NMC) 6-layer Primitive Equation Model and from the TDL Three-Dimensional Trajectory Model as predictors. One hundred and three predictors were used in the initial experiment consisting of dynamic and kinematic parameters, geopotential height and thickness, humidity, stability parameters, temperatures, winds, and miscellaneous parameters.

Initially the predictand was determined from nationally transmitted facsimile radar summary maps and finally from manually digitized radar (MDR) data and severe reports.

TDL has continued improving their equations over the years. J. P. Charba (1975) described TDL's short-range equation for severe local storms (2-6 hours after data observation time). The forecast probabilities of tornadoes, hail, and damaging wind obtained from this equation are transmitted to the NWS three times daily via teletype, for 90 nm x 135 nm (predictand) rectangular areas. Screening regression was also used to derive this equation. The predictors are derived mainly from observed data (not MOS).

## 6. Conclusion

The examples mentioned in section 5 are only a few of a great many attempts made by statisticians and meteorologists to simplify the task of dealing with a seemingly infinite number of variables. While the attempts mentioned are by no means all inclusive, they do show a variety of ways in which the screening procedure can be applied to the field of meteorology.

The screening procedure can be a useful procedure when a large number of predictors are considered and elimination of the insignificant ones is desired. There are, of course, shortcomings such as the equations' dependency on the sample used to develop it, the necessity to update the equations as the sampling data bases change or time steps change, the fact that the "best" set of variables may not be included in the final equation, and so on. However, as stated earlier in this paper, the screening procedure can help meteorologists confirm their notions of which predictors are most significant on the basis of theory.

## Chapter 4

### MULTIPLE DISCRIMINANT ANALYSIS

by Captain William W. Neubert

#### 1. Rationale for Multiple Discriminant Analysis vs Regression.

Impetus for the use of the multiple discriminant analysis (MDA) technique in establishing probabilities arises from a need to classify observed phenomenon into pre-established groups. MDA is well suited to performing this classification where the groups involved have no particular order or ranking.

Within a set of previously observed or recorded data, we might be interested in learning where a new observation might fit into this group. That is, where does our new observation lie with respect to the mean of the group? Such predictions could best be handled through multiple regression. Suppose, however, that we have several groups into which to classify our observations--no one group having any particular rank or order with respect to any other group. We might, for example be trying to generate the probability for the occurrence of a particular wind direction or a certain type of precipitation. In these two examples the classes into which we could separate the observed phenomena bear no ranking among them; i.e.,: snow is neither better or worse than rain; it's just different. The need is clear, however, for some method to predict into which of the unordered classes future observations will fall. MDA provides one way to derive the probability forecasts for such events. Multiple regression techniques would be more appropriate if we were dealing with ordered or scaled data such as temperature or wind velocity. In summary, multiple regression helps define distinctions within groups; whereas, discriminant analysis delineates distinctions between groups.

The discriminant function eliminates the need for looking at the measurements one at a time, only to find that the overlap of the data obscures any conclusions we might have been able to draw from the observations. MDA uses a set of weighted coefficients as multiples of several of the selected variables to produce a sum of products that is a single discriminant score. This score makes the best use of all the information contained in the variables we have selected to use. Given the groups involved, the computation develops the best set of weights possible from the measurements, and, in effect, sifts out the important differences that best separate the groups. The overlap in the raw data that had acted to obscure these differences is then reduced or removed. The technique is improved, or limited as the case may be, by the amount of information contained in the original predictor variables about the phenomena we are trying to predict. (Rulon, 1951, pp 82-3).

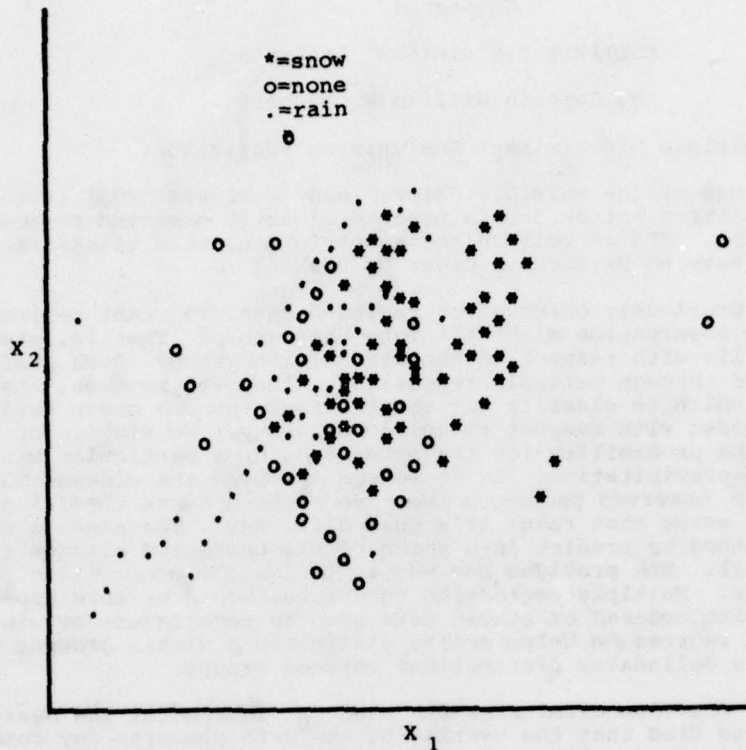
#### 2. Graphical Interpretation.

The fundamental idea of the discriminant function is best understood by viewing a graphical depiction of the results. Suppose two variables,  $X_1$  and  $X_2$  say, are considered meteorologically significant in the forecasting of precipitation. We wish to use the two variables to predict the occurrence of rain, or snow, or no precipitation at all. The forecast is to be valid six hours from the time of the observation of the values of  $X_1$  and  $X_2$ . Now, assume that we have accumulated a sample of three years of data on the two variables. We also know the type of weather that occurred six hours after the recording of the corresponding values of  $X_1$  and  $X_2$ . A graph can be constructed showing the values of  $X_1$  and  $X_2$  and the type of weather corresponding to each set of observations (Figure 1). The weather would be coded with appropriate symbology: (\*)=snow; (·)=rain; and (o)=none.

The object of this exercise would be to take newly observed values of our predictors and, using the graph, produce a forecast of the probability of rain or snow six hours hence. It is clear, however, that our graph of the raw predictors has produced a great deal of overlap among data points, and that such a prediction is all but impossible.

Now suppose we develop a particular computational technique that uses our accumulated data to produce two new values:

$$Y_1 = V_1X_1 + V_2X_2 \text{ and } Y_2 = V_3X_1 + V_4X_2.$$



NOTE: Figures 1 and 2 are exaggerated examples from Fictitious data, and are shown only to provide clarity of meaning.

Figure 1. Graph of Weather Predictand in raw predictor space .

The functions  $Y_1$  and  $Y_2$  are weighted sums of the original predictor variables and, as we shall see, serve as a better tool for prediction than the raw predictors alone. If we have carefully selected our predictors to include that set of variables that contain the most information about the phenomena we are trying to predict, then the plot of  $Y_1$  and  $Y_2$ , with corresponding weather, might look something like Figure 2.

These new weighted sum functions, called discriminants, provide a great deal more separation between the several distinct groups in our precipitation example. When compared to the plot of the raw predictor data, we can easily see how the discriminants "stretch" the data apart, reducing or eliminating the overlap. For clarity, these two sample figures have been greatly exaggerated, and, thus, present an oversimplified version of the results of a very complicated mathematical process. (Tanur, 1972, pp. 376-80).



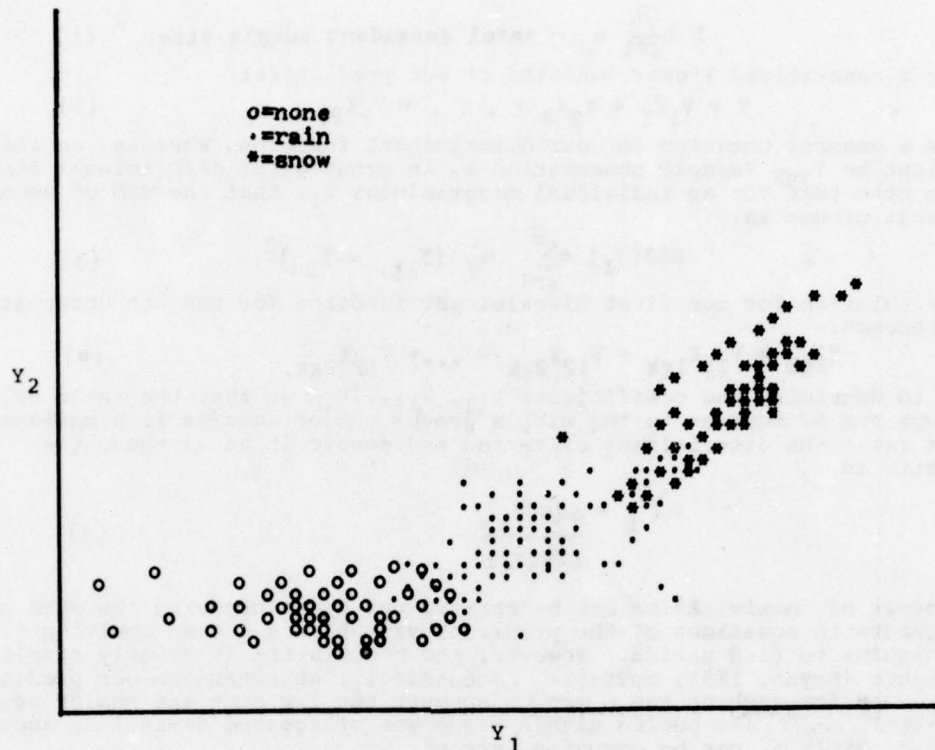


Figure 2. Graph of Weather Predictand in Discriminant Space.

### 3. Mathematical Procedure.

In line with the simplified example discussed above, when the number of group classifications exceeds two, the discriminant analysis becomes multi-dimensional. If there are  $G$  of these groups, then there are  $G$  points in space each representing one of the means of one of the  $G$  groups. If the number of predictor variables is at least  $G-1$ , then these group means will define a discriminant space of  $G-1$  dimensions (Bryan, 1951, p. 90).

The dimensions or directions delineate areas of major differences between the groups. In principle components analysis the concept of dimensions or directions in space is closely related to the algebraic idea of linear combinations. To study the directions of group differences, then, we are tasked with finding linear combinations of the original predictor variables that exhibit large differences in group means. Multiple discriminant analysis is a technique for finding those combinations that will separate the group means of the predictor variables to the maximum degree allowed by the predictor variables we have chosen (Tatsuoka, 1971, p. 157).

Assume that we have  $G$  mutually exclusive and exhaustive groups into which we are to classify future observations. We have chosen a group of predictors that contain information about which of the  $G$  groups future observations will fall into. We denote these predictors by  $X_p$ , numbering from  $p=1 \dots$  to a total of  $P$  variables. Assume we have a dependent sample of data compiled from observations of our  $X_p$  predictors, and let the number of observations in each group of this sample be  $n_g$  ( $g=1, \dots, G$ ), so that

$$N = \sum_{g=1}^G n_g = \text{total dependent sample size.} \quad (1)$$

Now consider a generalized linear function of our predictors:

$$Y = V_1 X_1 + V_2 X_2 + \dots + V_P X_P. \quad (2)$$

$Y$  represents a general notation for our discriminant function; whereas, an individual value might be  $Y_{jgk}$  (sample observation  $k$ , in group  $g$ , of discriminant function  $j$ ). We then note that for an individual discriminant  $Y_1$ , that the sum of squares between (among) groups is:

$$SSB(Y_1) = \sum_{g=1}^G n_g (Y_{1g.} - \bar{Y}_{1..})^2. \quad (3)$$

The complete notation for our first discriminant function for the  $k$ th observation of group  $g$  becomes:

$$Y_{1gk} = V_{11} X_{1gk} + V_{12} X_{2gk} + \dots + V_{1P} X_{Pgk}. \quad (4)$$

The task is to determine the coefficients  $V_{11}, V_{12}, \dots, V_{1P}$  so that the ratio of the between groups sum of squares to the within groups sum of squares is a maximum. We call this ratio the discriminant criterion and denote it as  $\lambda$ ; thus, the maximized ratio is

$$\lambda_1 = \frac{SSB(Y_1)}{SSW(Y_1)} \quad (5)$$

This process of maximization can be carried out by expressing the sums of squares as quadratic equations of the predictor variables and then applying differential calculus to find maxima. However, the computation is greatly simplified by using algebra (Bryan, 1951, pp90-95). Accordingly, we return to our predictor variables  $X_p$ , and for each of the  $G$  groups compute the raw sums and sum of squares for each  $X_p$  ( $p=1, \dots, P$ ). The pooled within group sum of squared deviations about the group mean,  $SSW(X_p)$ , can be computed, since:

$$\begin{aligned} SSW(X_p) &= \sum_{g=1}^G \sum_{k=1}^{n_g} (X_{pgk} - \bar{X}_{pg.})^2 \\ &= \sum_{g=1}^G \left[ \sum_{k=1}^{n_g} X_{pgk}^2 - \frac{(\sum_{k=1}^{n_g} X_{pgk})^2}{n_g} \right] \quad (6) \\ &\quad (p=1 \dots P). \end{aligned}$$

The initial calculation of the raw sum and sum of squares will yield the total sum of squared deviations about the grand mean,  $SST(X_p)$ , because:

$$\begin{aligned} SST(X_p) &= \sum_{g=1}^G \sum_{k=1}^{n_g} (X_{pgk} - \bar{X}_{p..})^2 \\ &= \sum_{g=1}^G \sum_{k=1}^{n_g} X_{pgk}^2 - \frac{(\sum_{g=1}^G \sum_{k=1}^{n_g} X_{pgk})^2}{N} \quad (7) \\ &\quad (p=1, \dots, P) \end{aligned}$$

Subtracting the pooled within group sum of squares,  $SSW(X_p)$ , from the total sum of squares,  $SST(X_p)$  yields the sum of squared deviations between group means and the grand mean. That is:

$$SSB(X_p) = SST(X_p) - SSW(X_p) \quad (p=1, \dots, P)$$

since

$$\begin{aligned}
 \text{SSB}(X_p) &= \sum_{g=1}^G \sum_{k=1}^{n_g} (\bar{x}_{pg.} - \bar{x}_{p..})^2 = \sum_{g=1}^G n_g (\bar{x}_{pg.} - \bar{x}_{p..})^2 \\
 &= \sum_{g=1}^G \left[ \frac{\left( \sum_{k=1}^{n_g} x_{pgk} \right)^2}{n_g} \right] - \frac{1}{N} \left[ \sum_{g=1}^G \sum_{k=1}^{n_g} x_{pgk} \right]^2 \quad (8) \\
 &\quad (p=1, \dots, P)
 \end{aligned}$$

Additionally, the sums of cross products between each of the raw predictors must be used to calculate the sum of products between and within groups for variable  $X_p$  and  $X_q$ , let's say, with  $p, q=1, \dots, P$  and  $p \neq q$ . The notation is  $\text{SPW}(X_p X_q)$  and  $\text{SPB}(X_p X_q)$ . For the within groups determination we have:

$$\begin{aligned}
 \text{SPW}(X_p X_q) &= \sum_{g=1}^G \sum_{k=1}^{n_g} (x_{pgk} - \bar{x}_{pg.})(x_{qgk} - \bar{x}_{qg.}) \\
 &= \sum_{g=1}^G \left[ \sum_{k=1}^{n_g} x_{pgk} x_{qgk} - \frac{\left( \sum_{k=1}^{n_g} x_{pgk} \right) \left( \sum_{k=1}^{n_g} x_{qgk} \right)}{n_g} \right] \quad (9) \\
 &\quad (p, q=1, \dots, P; p \neq q)
 \end{aligned}$$

Using the raw sums and sum of crossproducts we can obtain the total sum of products,  $\text{SPT}(X_p X_q)$ , since:

$$\begin{aligned}
 \text{SPT}(X_p X_q) &= \sum_{g=1}^G \sum_{k=1}^{n_g} (x_{pgk} - \bar{x}_{p..})(x_{qgk} - \bar{x}_{q..}) \\
 &= \sum_{g=1}^G \sum_{k=1}^{n_g} x_{pgk} x_{qgk} - \frac{\left( \sum_{g=1}^G \sum_{k=1}^{n_g} x_{pgk} \right) \left( \sum_{g=1}^G \sum_{k=1}^{n_g} x_{qgk} \right)}{\sum_{g=1}^G n_g} \quad (10) \\
 &\quad (p, q=1, \dots, P; p \neq q)
 \end{aligned}$$

As before, the sum of products between groups  $\text{SPB}(X_p X_q)$  is obtained by subtracting the sum of products within groups,  $\text{SPW}(X_p X_q)$ , from the total sum of products,  $\text{SPT}(X_p X_q)$ ; that is:

$$\text{SPB}(X_p X_q) = \text{SPT}(X_p X_q) - \text{SPW}(X_p X_q) \quad (p, q=1, \dots, P; p \neq q) \quad (11)$$



This results since:

$$\begin{aligned}
 SPB(X_p X_q) &= \sum_{g=1}^G \sum_{k=1}^{n_g} (\bar{X}_{pg.} - \bar{X}_{p..}) (\bar{X}_{qg.} - \bar{X}_{q..}) \\
 &= \sum_{g=1}^G n_g (\bar{X}_{pg.} - \bar{X}_{p..}) (\bar{X}_{qg.} - \bar{X}_{q..}) \quad (12) \\
 &= \sum_{g=1}^G \left[ \frac{\left( \sum_{k=1}^{n_g} x_{pgk} \cdot \sum_{k=1}^{n_g} x_{qgk} \right)}{n_g} - \frac{\left( \sum_{g=1}^G \sum_{k=1}^{n_g} x_{pgk} \right) \left( \sum_{g=1}^G \sum_{k=1}^{n_g} x_{qgk} \right)}{\sum_{g=1}^G n_g} \right] \\
 &\quad (p, q=1, \dots, P; p \neq q)
 \end{aligned}$$

A matrix W, representing the pooled within group matrix, is then constructed from these derived quantities:

$$\underline{W} = \begin{bmatrix} SSW(X_1) & SPW(X_1 X_2) & \dots & SPW(X_1 X_{P-1}) & SPW(X_1 X_P) \\ SPW(X_1 X_2) & SSW(X_2) & \dots & SPW(X_2 X_{P-1}) & SPW(X_2 X_P) \\ \vdots & \vdots & & \vdots & \vdots \\ \vdots & \vdots & & \vdots & \vdots \\ SPW(X_1 X_{P-1}) & SPW(X_2 X_{P-1}) & \dots & SSW(X_{P-1}) & SPW(X_{P-1} X_P) \\ SPW(X_1 X_P) & SPW(X_2 X_P) & \dots & SPW(X_{P-1} X_P) & SSW(X_P) \end{bmatrix} \quad (13)$$

The data previously accumulated from (12) about the sums of products between the various groups is then used to construct a pooled between groups matrix B.

$$\underline{B} = \begin{bmatrix} SSB(X_1) & SPB(X_1 X_2) & \dots & SPB(X_1 X_{P-1}) & SPB(X_1 X_P) \\ SPB(X_1 X_2) & SSB(X_2) & \dots & SPB(X_2 X_{P-1}) & SPB(X_2 X_P) \\ \vdots & \vdots & & \vdots & \vdots \\ \vdots & \vdots & & \vdots & \vdots \\ SPB(X_1 X_{P-1}) & SPB(X_2 X_{P-1}) & \dots & SSB(X_{P-1}) & SPB(X_{P-1} X_P) \\ SPB(X_1 X_P) & SPB(X_2 X_P) & \dots & SPB(X_{P-1} X_P) & SSB(X_P) \end{bmatrix} \quad (14)$$

The computational procedure for deriving the discriminant functions will make use of the fact that all the discriminants are calculated from these matrices W and B. The actual process involves first pre-multiplying B by the inverse of W, and then determining the eigenvalues and eigenvectors of the matrix that results. The latter operation produces the solution to the determinant equation

$$\begin{aligned}
 |\underline{W}^{-1} \underline{B} - \lambda \underline{I}| &= 0 \quad (15) \\
 (\underline{I} &= \text{the unit matrix})
 \end{aligned}$$

The eigenvectors are solutions to the equation:

$$\left[ \underline{W}^{-1} \underline{B} - \lambda_j \underline{I} \right] \underline{V}_j = 0 \quad (j=1, \dots, \min(G-1, P)) \quad (16)$$

The elements of the eigenvectors ( $\underline{V}_j$ ) represent the weights to be applied to the original X predictors in our linear function:

$$Y_j = V_{j1}X_1 + V_{j2}X_2 + \dots + V_{jP}X_P \quad (j=1, \dots, \min(G-1, P)) \quad (17)$$

The value(s) of  $\lambda_j$  are the roots of the characteristic equation resulting from the expansion of the determinant in (15); they represent the ratios of the corresponding  $Y_j$ 's sum of squares between and within groups, namely:

$$\lambda_j = \frac{\sum_{g=1}^G n_g (\bar{Y}_{jg} - \bar{Y}_{j..})^2}{\sum_{g=1}^G \sum_{k=1}^n (Y_{jgk} - \bar{Y}_{jg})^2} \quad (j=1, \dots, \min(G-1, P)) \quad (18)$$

(Miller, 1961, pp6-13)

The characteristic equation derived from  $\underline{W}^{-1} \underline{B}$  will no doubt have several roots. From each of these roots we can calculate an eigenvector  $\underline{V}_j$ , the elements of which represent a new set of combining weights...  $V_{j1}, V_{j2}, \dots, V_{jP}$ . If these new elements are used to form a second linear combination

$$Y_2 = V_{21}X_1 + V_{22}X_2 + \dots + V_{2P}X_P$$

the  $Y_2$  is a new discriminant function that is uncorrelated with  $Y_1$ , but having its ratio  $\lambda_2$  of sum squares between groups to sum of squares within groups a maximum after  $\lambda_1$ . As the process continues, each successive  $Y$  is linearly uncorrelated with any of the preceding linear combinations, and has its ratio of sum of squares between and within groups a maximum. The values created ( $Y_1, Y_2, \dots, Y_{\min(G-1, P)}$ ) are the first, second, ...etc. discriminant functions for optimally differentiating among the  $G$  groups. Thus, by having several discriminators, we are shown the dimensions of the differences between groups and the direction along which maximum group differences occur.

In principal component analysis the dimension corresponding to the first component has maximum variance; whereas, the second component's dimension has maximum variance among those uncorrelated with the first, and so on.

In MDA  $\lambda$ , the ratio of between to within groups sums of squares, merely takes the place of variance as the factor determining the various successive dimensions. It should be noted, however, that in the discriminant space the axes are not necessarily mutually orthogonal even though they are uncorrelated. Although the discriminant function performs a linear transformation on the original X predictor axes, the rotation that occurs may indeed be an oblique rotation (Tatsuoka, 1971 pp 162-3).

We can reiterate the entire procedure in a stepwise fashion:

- a. Assemble the within groups sum of squares from (6).

- b. Compute the total sum of squares about the grand mean,  $SST(X_p)$ , using (7).
- c. Next, assemble the data from steps a and b and calculate the sum of squared deviations between group means and the grand mean using:

$$SSB(X_p) = SST(X_p) - SSW(X_p).$$

- d. Compute the sum of products between groups using (10) and (12); and then construct the matrix  $B$  as in (14).
- e. Derive the matrix  $W^{-1}B$  whose eigenvalues and eigenvectors we wish to find, using (13) and (14).
- f. Develop the characteristic equation for the matrix found in e using the relation:

$$|W^{-1}B - \lambda I| = 0$$

in expanded form.

- g. Solve for the roots of the equation in step f.
- h. These roots,  $\lambda_1, \lambda_2, \dots$  etc., are then used to determine the eigenvectors  $V_1, V_2, \dots$  etc. whose elements are the weights to be applied to the original predictors in our linear relation for defining the various  $Y_j$ s. The procedure for finding these characteristic vectors, as outlined by Tatsuoaka follows:

(1) For each given eigenvalue  $\lambda_j$ , compute or form the matrix  $W^{-1}B - \lambda_j I$  by subtracting  $\lambda_j$  from each diagonal element of  $W^{-1}B$ .

(2) Compute the adjoint of the matrix  $W^{-1}B - \lambda_j I$ , ( $\text{adj } W^{-1}B - \lambda_j I$ ). The adjoint or adjugate of a matrix is found by gathering the cofactors of all the elements of the matrix and then using these cofactors to form a new matrix  $\text{adj}(A)$ . However, the new matrix is constructed so that the cofactor of the elements of the first row of the original matrix  $A$  are now used to form the elements of the first column of  $\text{adj}(A)$ --those of the second row of the cofactor matrix become the second column of the adj matrix, etc.

(3) Next divide the elements of any column of  $\text{adj } (W^{-1}B - \lambda_j I)$  by the square root of the sum of the squares of these elements. The resulting numbers are the elements of the eigenvector  $V_j$ .

i. Each value of  $\lambda_j$  produces an eigenvector  $V_j$  whose elements are the weights or coefficients in the linear equation for each of the discriminant functions  $Y_j$ . That is:

$$V_1 = \begin{bmatrix} v_{11} \\ v_{12} \\ v_{13} \end{bmatrix}, \quad \text{and} \quad V_2 = \begin{bmatrix} v_{21} \\ v_{22} \\ v_{23} \end{bmatrix}$$

are eigenvectors derived from discriminant criteria  $\lambda_1$  and  $\lambda_2$  (eigenvalues) to be used to set up the linear relations

$$Y_1 = v_{11}X_1 + v_{12}X_2 + v_{13}X_3$$

$$\text{and} \quad Y_2 = v_{21}X_1 + v_{22}X_2 + v_{23}X_3$$

respectively. (Tatsuoka, 1971, pp166-70).

#### 4. Selecting the Predictors.

In applications of multiple regression analysis it is generally agreed among researchers that most of whatever predictability is in a particular set of predictors is contained in a reasonably small subset of the total group. Indeed, the sheer unreasonableness of the computations involved in evaluating huge matrices constructed of all the predictors led to the development of a technique called screening regression. This method was first developed by Joseph G. Bryan (1951). Its purpose was to define a small group of the predictors that contained most of the predicting information required.



The technique determines which variable has the highest correlation with the predictand and chooses that variable as the first predictor. Of the remaining variables, the predictor having the highest partial correlation, after the effects of the first predictor are removed, is the next one selected. The process continues--each successive selection being based on the highest partial correlation after the effects of previously selected variables have been removed. The screening ceases at a pre-determined point when further significant improvement is not obtained. If the original total set of predictor variables was large, this method may provide a considerable savings in labor.

Such a stepwise screening process has also been developed for applications to MDA. Selection is made from a large set of variables to first find that variable which provides the greatest degree of discrimination between groups as measured by the generalized Mahalanobis distance,  $D^2$ . This is discussed later. This variable is our first predictor  $X^{(1)}$ . Among the remaining  $P-1$  variables a second is selected which together with the first gives the highest value of  $D^2$ . This second variable is denoted  $X^{(2)}$ . Added predictors continue to be selected in like manner--always yielding a maximum value of  $D^2$  when combined with the previous predictors. (Miller, 1961, pp34-36)

The extension of the Mahalanobis  $D^2$  to situations involving more than 2 groups was first developed by Rao (1952) and Bryan (1950). For  $P$  predictors the test statistic  $D_p^2$  is defined as

$$D_p^2 = (n - G \cdot P) \cdot \text{trace } \underline{W}^{-1} \underline{B} \quad (19)$$

$\underline{W}$  and  $\underline{B}$  are the matrices defined earlier and  $\text{trace } \underline{W}^{-1} \underline{B}$  refers to the sum of the diagonal elements of the matrix  $\underline{W}^{-1} \underline{B}$ .  $n$  is one less than the number of independent observations in the sample. The distribution of  $D_p^2$  is estimated, for  $n$  large, as

$$D_p^2 \sim \chi^2(P(G-1)). \quad (20)$$

A modification of this relation will be used to assist in determining if successively selected predictors are statistically significant.

The actual screening procedure is applied as follows. For every one of the  $P$  available predictors  $X_p$  ( $p=1, \dots, P$ ) determine the quantities  $SSW(X_p)$  and  $SSB(X_p)$ . Using these, find

$$\text{trace } \underline{W}^{-1} \underline{B}(X_p) = \frac{SSB(X_p)}{SSW(X_p)} \quad (p=1, \dots, P) \quad (21)$$

To select the first predictor  $X^{(1)}$ , where the total may be  $r$  ( $r \leq P$ ), we use the criterion

$$\text{trace } \underline{W}^{-1} \underline{B}(X^{(1)}) \geq \text{trace } \underline{W}^{-1} \underline{B}(X_p). \quad (p=1, \dots, P) \quad (22)$$

if  $X^{(1)}$  can be determined to be statistically significant. A description of this test is given later.

For the  $P-1$  remaining variables calculate  $SPW(X^{(1)}X_p)$ , and  $SPB(X^{(1)}X_p)$ . From these and the values of  $SSW(X_p)$  and  $SSB(X_p)$  already computed we then derive:

$$\text{trace } \underline{W}^{-1} \underline{B}(X^{(1)}X_p) = \text{trace} \begin{bmatrix} SSW(X^{(1)}) & SPW(X^{(1)}X_p) \\ SPW(X^{(1)}X_p) & SSW(X_p) \end{bmatrix}^{-1} \begin{bmatrix} SSB(X^{(1)}) & SPB(X^{(1)}X_p) \\ SPB(X^{(1)}X_p) & SSB(X_p) \end{bmatrix} \quad (23)$$

( $p=1, \dots, P; X_p \neq X^{(1)}$ )

The selection of the second predictor  $X^{(2)}$  is conditional on the fact that  $X^{(1)}$  has already been selected. The criterion for  $X^{(2)}$  is

$$\text{trace } W^{-1} B(X^{(1)} X^{(2)}) \geq \text{trace } W^{-1} B(X^{(1)} X_p) \quad (24)$$

$$(p=1, \dots, P; X_p \neq X^{(1)})$$

again provided that  $X^{(2)}$  is statistically significant. So then, a general form can be used to choose those to be selected out of the remaining predictors; this is:

$$\text{trace } W^{-1} B(X^{(1)} X^{(2)} \dots X^{(S-1)} X^{(S)}) \geq \text{trace } W^{-1} B(X^{(1)} X^{(2)} \dots X^{(S-1)} X_p)$$

$$(p=1, \dots, P; X_p \neq X^{(1)}, \dots, X^{(S-1)})$$

The entire procedure continues until  $r$  predictors have been chosen. The total number of selected predictors  $r$  is completely determined when the variable  $X^{(r+1)}$  fails to show statistical significance. (Miller, 1961, pp 43-47)

A particular predictor, say  $X^{(1)}$  is deemed significant if

$$D_1^2 - D_0^2 > X^2(\frac{\alpha^*}{P})(G-1). \quad (25)$$

That is, the criterion for deciding when to discontinue selection is based on Rao's test on  $D^2$ , with a modification introduced to assist in determining the significance of a newly selected variable.

After choosing a predictor from a set of variables, a chi square test is performed and the critical value of  $\alpha^*$  is set somewhat arbitrarily at .05. This allows a 1/20 chance of the predictor chosen being significant when in fact it is not. The test size is then designated as  $\alpha = .05$ . In our selection procedure, however, a predictor variable is chosen out of the  $P$  total because it maximizes some function of the test statistic. It is necessary, therefore, to consider at what level of probability the critical value of  $\alpha^*$  should be set while still allowing for only a 1/20 chance of error.

Let  $\alpha^*$  represent the probability that one or more of these predictors are adjudged significant when, in reality, not one of the  $P$  total is significant. So,  $1 - \alpha^* = (1 - \alpha)^P$ , provided that  $P$  tests are independent. If  $\alpha$  is small, we assume that  $(1 - \alpha)^P \text{ approx. } = (1 - P\alpha)$ . Then,  $(1 - \alpha^*) \text{ approx. } = (1 - P\alpha)$ , and then

$$\alpha \text{ approx. } = \frac{\alpha^*}{P} \quad (26)$$

Let  $\chi_{\alpha^*}^{2*}$  be the critical chi square value when selection is performed, and

$$\chi_{\alpha^*}^{2*} \equiv \chi_{\alpha}^2. \quad (27)$$

Then, if  $\alpha^*$  denotes the desired size of the selection test, then

$$\chi_{\alpha^*}^{2*} \text{ approx. } = \chi^2 \frac{\alpha^*}{P}$$

$P$ , again, is the total number of predictors possible. Therefore, the critical value of  $\chi_{\alpha^*}^{2*}$  for testing the significance of the  $S^{\text{th}}$  selected predictor  $X^{(S)}$  is:

$$\chi_{\alpha^*}^{2*} = \chi^2 \left( \frac{\alpha^*}{P-S+1} \right) \quad (28)$$

$$(S=1, \dots, P)$$

Further, based on maximizing  $D^2$  as our selection criterion and a level of significance as expressed in (28), the test of the  $S^{\text{th}}$  selected predictor  $X^{(S)}$  is:

$$(D_S^2 - D_{S-1}^2) > \chi^2 \left( \frac{\alpha^*}{P-S+1} \right) (G-1) \quad (29)$$

( $S=1, \dots, P$ )  
(Miller, 1961, pp.49-51)

## 5. Estimating Probabilities.

The application of MDA to meteorological parameters seems particularly well suited to predicting the occurrence of events that can be classified into unordered groups. Miller's original work demonstrated that problems can arise in achieving adequate discrimination, if multivariate normality is not present. The original idea had been to select predictors and obtain *a posteriori* probabilities using Bayes' theorem--assuming this normality and equal dispersion. It soon became apparent that these assumptions were not always tenable.

The first effort that was successful in using more than just the first discriminant function was made when a rectangular area was constructed around a new observation ( $Y_1Y_2$ ) in the two dimensional discriminant space. Group relative frequencies inside this area, constructed of observations of the development sample, were taken as estimates of the conditional probability distribution. A method was then employed that used the idea of the Euclidean distance as a way of defining a more desirable spherical area (neighborhood) about ( $Y_1Y_2$ ). This method turned out to be highly successful at providing valid probabilities in the multi-dimensional space.

The procedure was first developed by Fix and Hodges (1951) and requires computing the Euclidean distance  $D$  between the new observation ( $y$ ) and each of the observations of the dependent sample. So, for all the  $N$  observations, determine the weighted distances

$$D(y', y'_{gk}) = \left[ \frac{\lambda_1}{\lambda_1} (y'_1 - y'_{1gk})^2 + \dots + \frac{\lambda_j}{\lambda_1} (y'_j - y'_{jgk})^2 + \dots + \frac{\lambda_t}{\lambda_1} (y'_t - y'_{tgk})^2 \right]^{1/2} \quad (30)$$

$k=1, \dots, n_g$   
 $g=1, \dots, G$

where

$$y'_j = \frac{y_j - \bar{y}_{j..}}{\hat{\sigma}_{Y_j}}, \quad y'_{jgk} = \frac{y_{jgk} - \bar{y}_{j..}}{\hat{\sigma}_{Y_j}} \quad (31)$$

$k=1, \dots, n_g$   
 $g=1, \dots, G$   
 $j=1, \dots, t$

and into which we have substituted  $\bar{y}_{j..}$ , where

$$\bar{y}_{j..} = \sum_{g=1}^G \sum_{k=1}^{n_g} y_{jgk} / N$$

$$\text{and } \hat{\sigma}_{Y_j} = \left[ \frac{\sum_{g=1}^G \sum_{k=1}^{n_g} y_{jgk}^2 - \frac{(\sum_{g=1}^G \sum_{k=1}^{n_g} y_{jgk})^2}{N}}{N-1} \right]^{1/2} \quad (32)$$

$(j=1, \dots, t)$



The weights  $\lambda_j$  ( $j=1, \dots, t$ ) are the eigenvalues previously found. Here, they serve as a means of producing an arbitrary metric in the discriminant space which accounts for the relative importance of the discriminant functions. The metric in the discriminant space is chosen so that each dimension  $j$  has zero mean and variance equal to  $\lambda_j / \lambda_1$  ( $j=2, \dots, t$ ), and the first discriminant function is transformed to have unit variance.

The  $N$  distances  $D(y', Y'_{gk})$ , ( $k=1, \dots, n_g$ ;  $g=1, \dots, G$ ), are then ordered. The point whose distances to  $y$  is the least is ranked as first, the next closest is ranked second, etc., down to the point having the greatest distance. Next, a spherical area is drawn around the point  $y$  by choosing the  $h$  ( $h \leq N$ ) closest points. Associated with each of these  $h$  points is a particular group which occurred subsequent to the observation of that point. The ratio of the observed frequencies in each of the groups to the total number  $h$  determines estimates of the desired probabilities  $\hat{p}_g$ , ( $g=1, \dots, G$ )

## 6. Applications to Precipitation Forecasting.

What follows next is a brief summary of one aspect of the research originally performed by Miller in his 1961 work with the Hartford, Connecticut data. The example outlined was the more successful of the two projects undertaken, in that the results show a clearer discrimination between both amounts and types of precipitation. Some of the graphs and tables presented are taken directly from the original work with the author's permission.

Discriminant analysis seems particularly applicable to the problem of forecasting the type and amount of precipitation that will occur at a particular location at a specific time. The first step is to define the operationally significant conditions, i.e., decide what phenomena we wish to predict, and how much of each will be considered significant. Then divide or categorize the various phenomenon (or degree of phenomenon occurring) into distinct groups. For economy of effort, if the number of predictor variables is large, the screening technique shown in section 4 may be applied.

For this particular example a set of five conditions, describing various types and degrees of precipitation, were chosen as being operationally of interest. (see Table 1). The forecasts are to be valid six hours after the daily observations (predictors) were noted. The data base consisted of an independent sample of 221 observations over a 1 year period, and a dependent sample of 1096 observations over a prior 3 year period. A total of seven meteorological parameters noted at each of twenty-five stations in the United States were chosen as representing the total number of predictor variables possible in the sample. From these one-hundred and seventy-five predictors a group of sixteen were shown to possess significant information as determined from measures of the generalized distance  $D^2$ . Table 2 depicts the meteorological parameters in the total sample  $P$ ; whereas, Table 3 shows the final sixteen selected along with the calculated values for trace  $W^{-1}B$ ,  $D_S^2 - D_{S-1}^2$ , and  $X_{.05}^{2*}$ . Miller points out that the physical significance of just why one particular variable has greater utility in predicting the weather over any other variable would be very difficult to determine. Table 4 shows the characteristic roots and vectors for  $W^{-1}B(X^{(1)} \dots X^{(16)})$ , and it is interesting to note that in this example all four roots show statistical significance.

Each individual observation from the dependent and independent samples was plotted in two dimensions with one axis  $Y_1$  and the other  $(\lambda_2/\lambda_1)^{1/2} Y_2$ . This is according to the non-parametric procedure outlined in section 5. The construction of the dimensions is done so as to enable a circle to correspond to the area (neighborhood) described for the above mentioned non-parametric procedure for estimating conditional group probabilities. In Figure 3 we see a composite of all groups displayed, showing the fifty percent contour ellipses. Bivariate normality is assumed without, necessarily, having equal dispersions. Figure 4 shows the distributions of the dependent and independent sample points for group one only. The fifty percent contour ellipse for group one is projected onto this plot to show the degree of bivariate normality obtained. These latter two graphical depictions represent data for group one of the precipitation sample only. The remaining groups have similar, though somewhat less dense, distributions; and they are not shown for the sake of brevity. From these two and the remaining data plots Miller was able to conclude the following:

Table 1. DESCRIPTION OF THE PRECIPITATION GROUPS  
FOR THE HARTFORD, CONNETICUT EXAMPLE.

GROUP NUMBER	CONDITIONS
1	No precipitation of any kind over the forecast period.
2	Rain or freezing rain reported at some time over the forecast period in the amount of at least a trace but not more than .05 inches. No snow or sleet reported at any time over this period.
3	Snow or sleet reported at some time during the forecast period in the amount of at least a trace but not more than .05 inches of melted water equivalent.
4	Rain or freezing rain reported at some time during the forecast period in the amount of greater than .05 inches. No snow or sleet reported at any time over this interval.
5	Snow or sleet reported at some time over the forecast period in the amount of greater than .05 inches of melted water equivalent.

Table 2 Available Meteorological Predictors.

ELEMENT	NOTATION
Sea Level Pressure	(P)
Past 3 hour change in sea level pressure	$\Delta P$
Dry bulb temperature	T
Temperature-dew point depression	$T-T_d$
East-West wind component	u
North-South wind component	v
Total cloud cover	(N)

Table 2a. Specifications for the Precipitation Study.

SPECIFICATION	NOTATION	NUMERICAL VALUE
Number of groups	G	5
Observations in Group 1	$n_1$	817
Observations in Group 2	$n_2$	135
Observations in Group 3	$n_3$	29
Observations in Group 4	$n_4$	92
Observations in Group 5	$n_5$	23
Total dependent sample size	N	1096
One less than the number of independent observations in dependent sample	n	1095
Total independent sample size	M	221
Number of available predictors	P	175
Forecast interval(hrs.)	H	0-6



Table 3 THE SELECTED PREDICTORS

SELECTED VARIABLE $x^{(s)}$	STATION	ELEMENT	Trace $\underline{W}^1 \underline{B}$	$D_S^2 - D_{S-1}^2$	$\chi_{.05}^{2*}$
$x^{(1)}$	Boston, Massachusetts	N	0.267	291.30	21.68 ↓
$x^{(2)}$	Portland, Maine	$\triangle P$	0.422	168.33	
$x^{(3)}$	St. Ste. Marie, Mich.	T	0.529	115.67	
$x^{(4)}$	Hartford, Connecticut	$T - T_d$	0.630	108.68	
$x^{(5)}$	Buffalo, New York	T	0.737	114.60	
$x^{(6)}$	Boston, Massachusetts	u	0.806	73.55	
$x^{(7)}$	Hatteras, N.C.	v	0.857	54.11	
$x^{(8)}$	Norfolk, Virginia	$\triangle P$	0.915	61.25	
$x^{(9)}$	New York, New York	T	0.962	49.40	
$x^{(10)}$	Portland, Maine	v	1.008	48.12	
$x^{(11)}$	Nantucket, Mass.	v	1.053	46.85	
$x^{(12)}$	Norfolk, Virginia	T	1.090	38.33	
$x^{(13)}$	Oklahoma City, Okla.	v	1.120	30.93	
$x^{(14)}$	Caribou, Maine	T	1.147	27.70	
$x^{(15)}$	Boston, Massachusetts	T	1.184	37.78	
$x^{(16)}$	Albany, New York	v	1.211	27.43	

Table 4  
CHARACTERISTIC ROOTS AND VECTORS OF  $W^{-1}B$  ( $x^{(1)} \dots x^{(16)}$ )  
FOR PRECIPITATION EXAMPLE

$s \backslash j$	$v_1(s)$	$v_2(s)$	$v_3(s)$	$v_4(s)$
1	-30.093	-8.343	-59.516	10.390
2	15.336	-3.952	-8.814	1.596
3	-0.043	-0.458	2.380	0.313
4	-4.385	-0.449	-1.713	0.342
5	2.310	-0.014	-1.287	-0.226
6	1.078	-0.495	-3.480	0.472
7	-1.533	-0.053	0.185	0.259
8	-9.788	1.076	-0.494	0.218
9	3.247	-0.474	0.311	-0.662
10	1.677	-1.052	1.316	-0.349
11	-1.881	-0.727	0.878	-0.702
12	-1.579	-0.115	0.686	0.387
13	-0.637	0.156	-1.495	0.041
14	-0.352	-1.023	-3.242	-0.496
15	1.709	2.125	4.911	0.661
16	1.000	1.000	1.000	1.000
Roots ( $\lambda_j$ )	0.925	0.158	0.091	0.037
$1084 \ln(1 + \lambda_j)$	710.0	158.3	94.3	39.1
$x_{.05}^{2*}(21 - 2 \cdot j)$	47.9	44.8	41.6	38.3

a. For the population distributions within each of the five groups the assumptions of bivariate normality and equal dispersions appear to be appropriate. Since the selection criterion  $D^2$  has optimum properties under such conditions, we can also assume that the predictors selected are those that contain the most information for predictive purposes.

b. There appears to be good agreement between the dependent (dots) and independent (crosses) sample observations. (Figure 4)

The data from the one, two, three, and four additional discriminant space was then used to develop probability predictions. Tables 5 and 6 are for the dependent sample only and are given just as the original data that were derived.

These tables are arranged in groups of five for each probability from left to right and the various meteorological groups are arranged vertically. The tables include the number of forecasts made within each group (F); the number of actual occurrences of the specific group when the predicted probability was in that range (U); the sum of the product of the probability for a forecast (P) and the number of forecasts (F) (FP); the sum of the product of P and (1 - P) for the forecasts (FP(1-P)); and lastly the computed values for  $\chi^2$ .

From these two and the remaining tables (not shown) the following:

a. The  $\chi^2$  tests for validity on the group probabilities show that there is no general tendency for the probabilities to become less valid with successive use of additional discriminant functions--this despite a general increase in the  $\chi^2$  values.

b. The simplified picture shown in Figure 3 disguises the fact that there is less tendency for the probabilities to be sharpened for conditions of snow, when the second discriminant function  $Y_2$  is used.

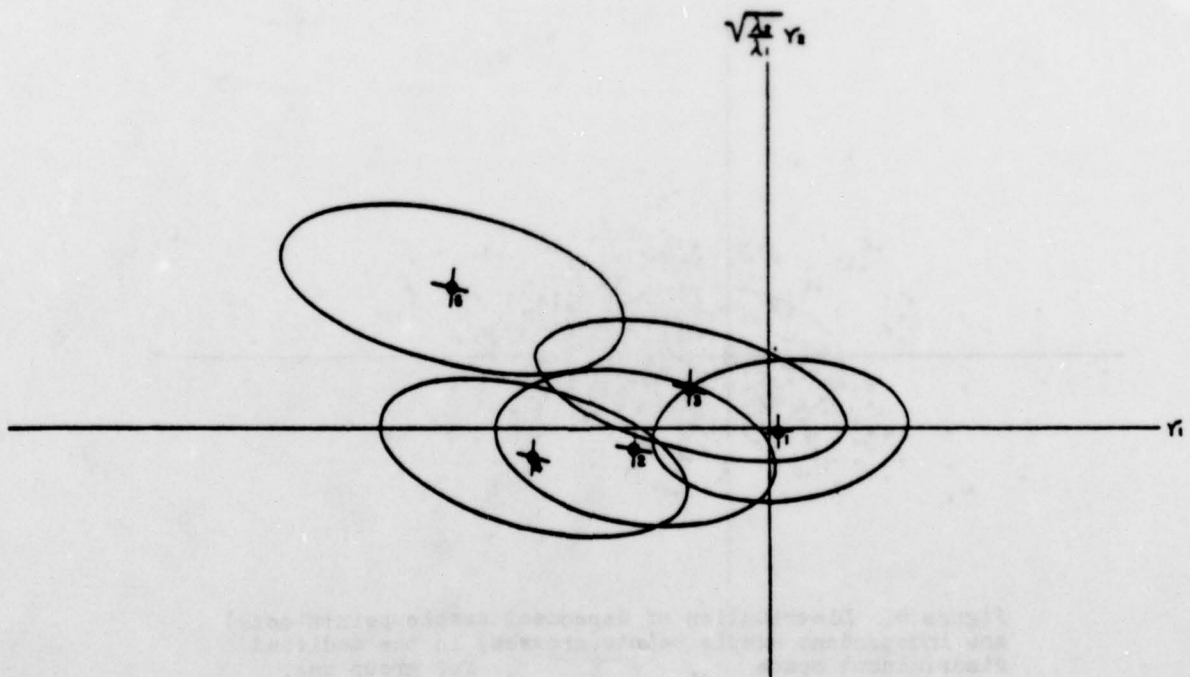


Figure 3. Mean and fifty per cent contour ellipse, assuming bivariate normality, in the modified discriminant space  $Y'_1, \sqrt{\frac{\lambda_2}{\lambda_1}} Y'_2$  for each group of the precipitation example.



In addition, it was also found that a discrepancy exists between the number of discriminant functions considered significant and the number actually producing the best results in the independent data. In the independent data inclusion of a third discriminant function produces marked deterioration in the accuracies of the predicted probabilities; even though, all four of the discriminant functions tested were found to be significant. (Miller, 1961, pp 95-130).

We can also draw a few more general conclusions from the data presented. Note from Figure 3 that the vertical axis seems to "discriminate" groups 3 and 5(snow) from groups 2 and 4(rain). Whereas, the horizontal axis seems to orient the groups by degree of precipitation--with the lesser amounts to the right and the greater amounts to the left. The success of the method can be measured to some extent by looking at the most general classifications possible--the prediction of precipitation or no precipitation. If we use the case where the probability of no precipitation was 0.50 or greater as our standard for categorically predicting no precipitation, the forecasts of no precipitation numbered 164 out of 221, with 144 correct. Precipitation was forecast 57 out of 221 times with 46 correct. In the independent data sample there was an overall percentage of 86% correct forecasts--in good agreement with the dependent sample of 87%.

The effectiveness of the multiple discriminant analysis technique in producing valid forecasts is demonstrated. The usefulness of MDA as a forecasting tool is, however, limited by resources at hand. The computations are all but impossible and highly impractical without the aid of a high speed computer. Further, other statistical methods, requiring much less labor, have recently come to light; and these procedures may in the long run totally replace MDA as a predictive method. (Tanur, 1972, pp 383-384)

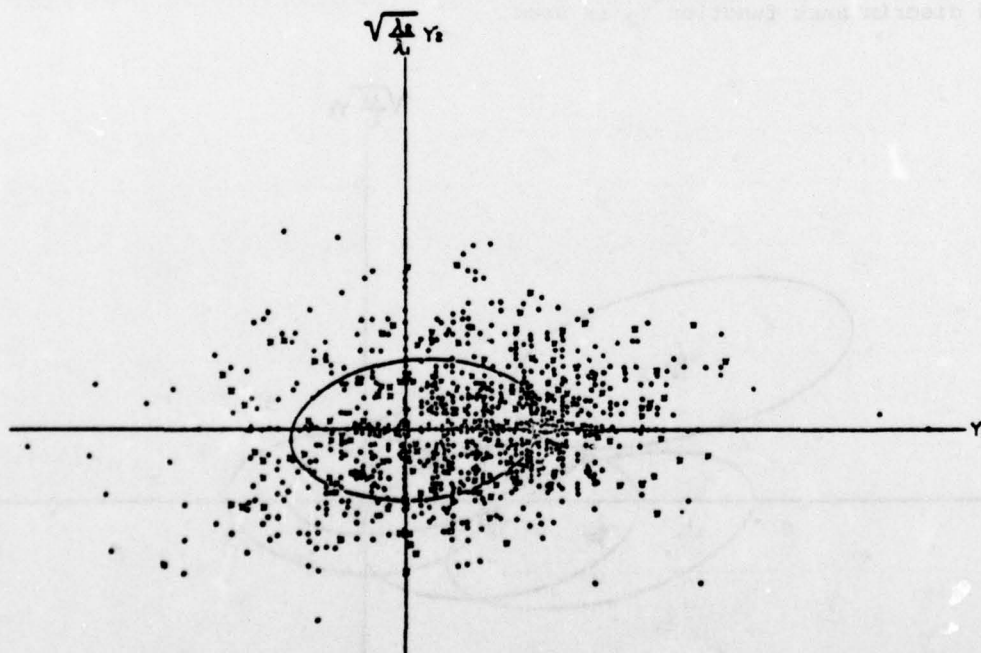


Figure 4. Distribution of dependent sample points(dots) and independent sample points(crosses) in the modified discriminant space

$$Y_1' \sqrt{\frac{\lambda_2}{\lambda_1}} Y_2'$$

for group one.

Table 5. SUMMARY OF PROBABILITY PREDICTIONS FOR PRECIPITATION EXAMPLE USING  $Y_1$  (DEPENDENT SAMPLE)

Group	$.0 \leq \hat{P} < .1$	$.1 \leq \hat{P} < .2$	$.2 \leq \hat{P} < .3$	$.3 \leq \hat{P} < .4$	$.4 \leq \hat{P} < .5$	$.5 \leq \hat{P} < .6$	$.6 \leq \hat{P} < .7$	$.7 \leq \hat{P} < .8$	$.8 \leq \hat{P} < .9$	$.9 \leq \hat{P} \leq 1.0$	Total	$\chi^2$
1 F U $\sum \hat{P}$ $\sum \hat{P}(1-\hat{P})$	63 5 4.56 4.21	77 11 12.72 10.52	56 15 15.12 11.02	13 5 4.92 3.04	35 19 16.00 8.67	41 21 22.84 10.07	42 25 27.88 9.36	103 81 79.00 18.32	93 82 79.64 11.40	573 553 555.96 15.92	1096 817 818.64 102.53	0.03
2 F U $\sum \hat{P}$ $\sum \hat{P}(1-\hat{P})$	645 17 15.76 14.80	180 28 29.44 24.45	119 31 31.08 22.92	110 38 39.20 25.09	42 21 19.04 10.39	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	1096 135 134.52 97.65	0.00
3 F U $\sum \hat{P}$ $\sum \hat{P}(1-\hat{P})$	1032 20 19.40 18.26	64 9 8 6.99	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	1096 29 27.40 25.25	0.10
4 F U $\sum \hat{P}$ $\sum \hat{P}(1-\hat{P})$	856 13 13.60 12.72	44 9 7.76 6.35	89 19 23.72 17.37	44 16 15.00 9.85	18 13 8.52 4.48	45 22 25.16 11.04	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	1096 92 93.76 61.81	0.05
5 F U $\sum \hat{P}$ $\sum \hat{P}(1-\hat{P})$	999 4 5.44 5.08	94 17 15.52 12.86	3 2 0.72 0.55	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	1096 23 21.68 18.49	0.09

Table 6. SUMMARY OF PROBABILITY PREDICTIONS FOR PRECIPITATION EXAMPLE USING  $Y_1$  AND  $Y_2$  (DEPENDENT SAMPLE)

Group		$.0 \leq \hat{P} < .1$	$.1 \leq \hat{P} < .2$	$.2 \leq \hat{P} < .3$	$.3 \leq \hat{P} < .4$	$.4 \leq \hat{P} < .5$	$.5 \leq \hat{P} < .6$	$.6 \leq \hat{P} < .7$	$.7 \leq \hat{P} < .8$	$.8 \leq \hat{P} < .9$	$.9 \leq \hat{P} \leq 1.0$	Total	$\chi^2$
1	F U $\sum \hat{P}$ $\sum \hat{P}(1-\hat{P})$	50 2 2.36 2.20	80 10 12.88 10.72	37 8 9.60 7.09	58 25 21.36 13.44	37 16 16.84 9.16	42 23 23.12 10.34	22 14 14.60 4.90	82 68 63.48 14.25	145 122 125.76 16.64	543 529 528.28 13.78	1096 817 818.28 102.52	0.02
2	F U $\sum \hat{P}$ $\sum \hat{P}(1-\hat{P})$	616 11 12.88 12.07	212 30 33.24 27.80	89 22 23.00 17.02	130 45 46.84 29.82	30 17 13.64 7.43	18 10 9.76 4.45	1 0 0.64 0.23	0 0 0 0	0 0 0 0	0 0 0 0	1096 135 140.00 98.82	0.13
3	F U $\sum \hat{P}$ $\sum \hat{P}(1-\hat{P})$	1034 19 17.56 16.62	46 6 7.04 5.92	15 4 3.92 2.89	1 0 0.32 0.22	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	1096 29 28.84 25.65	0.00
4	F U $\sum \hat{P}$ $\sum \hat{P}(1-\hat{P})$	821 8 6.84 6.40	100 11 15.52 13.00	54 16 13.88 10.29	63 20 22.16 14.31	20 11 9.20 4.96	20 13 11.32 4.90	12 10 8.00 2.66	6 3 4.32 1.21	0 0 0 0	0 0 0 0	1096 92 91.24 57.73	0.01
5	F U $\sum \hat{P}$ $\sum \hat{P}(1-\hat{P})$	1031 4 3.28 3.08	38 5 5.56 4.72	12 2 3.08 2.28	12 9 4.32 2.75	3 3 1.40 0.75	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	1096 23 17.64 13.58	2.12

## CHAPTER 5

### REGRESSION ESTIMATION OF EVENT PROBABILITIES

by CAPTAIN WENDELL POOL, JR.

This chapter discusses an application of Regression Estimation of Event Probabilities (REEP) by Bryan and Singer (1965). Their project was a statistical approach to predicting the probability of a first-term Navy recruit's reenlistment.

REEP was demonstrated as a useful prediction technique by Miller, Johnson, and Sorenson, see Miller (1964). Its development resulted from efforts to make multiple discriminant analysis more efficient.

Bryan and Singer studied the following problem:

Using available records on eligible first term electronics men (e.g., age, education, test scores, length of military duty, recruiting area) what sort of index or statistical digest of this information can be devised so as best to distinguish eventual reenlistees from non-reenlistees?

Their approach was to identify the independent components of significant data bearing on the question, then to develop a prediction formula into which an individual's particular data could be inserted. These data, termed variables, were selected from the expansive quantity of information compiled on each enlistee at enlistment. The first phase of the Navy study involved a partial screening of the possible significant variables, and eliminated many as not bearing on the question -- e.g., height, weight, color of hair. In the second phase, that in which Bryan and Singer were involved, the variables were more carefully screened using a multiple correlation approach. Reenlistment rates were calculated for one variable, while holding one or more other variables constant.

A variable that is the object of prediction or estimation will be called the predictand (in the present application, reenlistment), and the variables used to arrive at the prediction or estimation will be called predictors (for example, age, education, test scores). The type of prediction problem under consideration is that in which the predictand can assume any one of several distinct values, levels, or states (in the application to reenlistment prediction there are two states, to reenlist or not); and the object is to make use of the information available in the predictors to estimate the respective probabilities associated with each possible predictand state; that is, to estimate the chances that any specified state will be the one that the predictand actually assumes in a given instance.

Let the number of distinct states of the predictand be denoted by  $G$ . Unless otherwise noted, these  $G$  states (the case in point has two states of  $G$ , to reenlist or not) are exhaustive and mutually exclusive. If the predictors uniquely determine the predictand, the probability is unity for some one predictand state, as fixed by the predictors, and zero for all others. In a real situation, however, the predictors merely influence the probability by tending to favor the occurrences of some states more than others, depending on the given values of the predictors, and how the probability of occurrence is distributed over all  $G$  states. The statistical problem is to describe this distribution in terms of the predictors.

REEP uses multiple regression analysis. A dummy variable  $D_g$  ( $g=1,2,\dots,G$ ) is associated with each state,  $g$ , of the predictand:  $D_g = 1$  if state  $g$  occurs;  $D_g = 0$  if state  $g$  does not occur. Each dummy variable  $D_g$ , in turn, is treated as a predictand, to be estimated by a separate regression function (one for each dummy predictand). The device of using a common set of predictors for all  $D_g$ , as REEP does, insures that the sum of the estimated probabilities will be identically equal to one in every instance.

In the strict definition of the term, a regression function defines the conditional mean value of a predictand for any specified set of values of the predictors. The true conditional mean value of a dummy variable,  $D_g$ , is identically equal to the relative frequency -- hence, the conditional probability -- of the



occurrence of state,  $g$ , under the conditions defined by the predictors. If the exact mathematical specification of the regression function could be given, the true conditional probabilities could be determined from it.

In actuality, the mathematical specification of the regression function is not available. As a serviceable approximation to that function, REEP uses a linear expansion in terms of dummy variables, constructed from the predictors. These dummy variables can represent simple classes pertaining to individual variables, or, if desired, compound classes made of combinations of two or more variables.

In the foregoing, the term "regression function" will be applied to the expansion employed in REEP. The regression function  $\hat{D}_g$  for  $D_g$  is of the form:

$$\hat{D}_g = A_{0g} + A_{1g}x_1 + A_{2g}x_2 + \dots + A_{Mg}x_M \quad (g = 1, \dots, G) \quad (5-1)$$

The predictors  $x_1, x_2, \dots, x_M$  are selected by screening (see Chapter 3). The base constant,  $A_{0g}$ , and the coefficients  $A_{1g}, \dots, A_{Mg}$  are determined by least squares, so as to minimize the average value for the squared discrepancy  $(D_g - \hat{D}_g)^2$ .

$$\sum_{g=1}^G \hat{D}_g = 1 \quad (5-2)$$

Following is an analytical proof of (5-2).

Expanding Equation (5-1):

$$\begin{aligned} \hat{D}_1 &= A_{01} + A_{11}x_1 + A_{21}x_2 + \dots + A_{M1}x_M \\ \hat{D}_2 &= A_{02} + A_{12}x_1 + A_{22}x_2 + \dots + A_{M2}x_M \\ &\vdots \\ \hat{D}_G &= A_{0G} + A_{1G}x_1 + A_{2G}x_2 + \dots + A_{MG}x_M \end{aligned} \quad (5-3)$$

We shall prove that  $A_{01} + A_{02} + \dots + A_{0G} = 1$  and that  $A_{m1} + A_{m2} + \dots + A_{mG} = 0$  for  $m = 1, 2, \dots, M$ , where  $M$  is the number of selected predictors. This is sufficient to prove that  $\hat{D}_1 + \hat{D}_2 + \dots + \hat{D}_G = 1$ .

The matrix equation for generating the regression coefficients in the  $g$ -th equation in (5-3) is:

$$A_g = C^{-1}x' \hat{D}_g \quad (5-4)$$

in which the separate terms are defined as follows:

$A_g$  is a column vector with  $M + 1$  elements,  $(A_{0g} \ A_{1g} \ \dots \ A_{Mg})$

$C$  is a square matrix of order  $M + 1$  consisting of sums, sums of squares, and sums of cross-products of the predictor variables,

$$C = \begin{bmatrix} N & \sum x_1 & \sum x_2 & \dots & \sum x_M \\ \sum x_1 & \sum x_1^2 & \sum x_1 x_2 & \dots & \sum x_1 x_M \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum x_M & \sum x_M x_1 & \sum x_M x_2 & \dots & \sum x_M^2 \end{bmatrix} \quad (5-5)$$

All summations are from 1 to N, where N is the number of cases in the sample.

$X'$  is an  $M + 1$  by  $N$  matrix consisting of the individual values of the predictor variables,

$$X' = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{12} & \dots & x_{1N} \\ x_{21} & x_{22} & \dots & x_{2N} \\ \vdots & \vdots & & \vdots \\ x_{M1} & x_{M2} & \dots & x_{MN} \end{bmatrix} \quad (5-6)$$

$D_g$  is a column vector with  $N$  elements consisting of the individual values of the  $g$ -th dummy predictand,

$$D_g = (D_{1g} \ D_{2g} \ \dots \ D_{Ng}) \quad (5-7)$$

Equation (5-4) expresses a single set of regression coefficients. The matrix equation for all  $G$  sets of coefficients is

$$A = C^{-1}X'D \quad (5-8)$$

where  $A$  is an  $M + 1$  by  $G$  matrix consisting of the  $G$  column vectors  $A_1, A_2, \dots, A_G$ . Similarly,  $D$  is an  $N$  by  $G$  matrix consisting of the  $G$  column vectors  $D_1, D_2, \dots, D_G$ .

Define a column vector,  $e$ , consisting of  $G$  elements, each of which is unity:  $e = (1 \ 1 \ \dots \ 1)$ . Post-multiplying both sides of equation (5-8) by  $e$  gives

$$Ae = C^{-1}X'De \quad (5-9)$$

Note that  $Ae$  gives the sums that we require. That is, the  $m$ -th element of  $Ae$  is  $A_{m1} + A_{m2} + \dots + A_{mG}$ , and  $m$  ranges from zero through  $M$ .

Consider now the right-hand side of equation (5-9).  $De$  is a column vector with  $N$  elements, of which the  $n$ -th element is:

$$D_{n1} + D_{n2} + \dots + D_{nG}$$

This sum is identically equal to unity for all  $n$ , because one and only one of the  $G$  states ( $g$ ) must occur.  $D_g$  takes on the value 1, while remaining  $D$ 's are equal to zero.

Consider  $X'De$ . This is a column vector with  $M + 1$  elements:

$$(N \ \Sigma x_1 \ \Sigma x_2 \ \dots \ \Sigma x_M)$$

This is precisely the first column of the matrix  $C$ . Therefore

$$C^{-1}X'De = (1 \ 0 \ 0 \ \dots \ 0) \quad (5-10)$$

by the definition of the inverse of a matrix. Referring to equation (5-9), we see that

$$\begin{aligned} A_{01} + A_{02} + \dots + A_{0G} &= 1 \\ A_{m1} + A_{m2} + \dots + A_{mG} &= 0 \quad (m = 1, 2, \dots, M) \end{aligned} \quad (5-11)$$

which was to be proved!

Let us now return to the problem of the reenlistees. The next task is to select the variables.

The first step is to break the ordinary variables into groups of dummy variables. For example, age is broken into various (q) tentative dummy variables,  $T_q$ , such that:

$T_1 = 1$  if age is less than 17 years  
 $T_2 = 1$  if age 17 - 20\*  
 $T_3 = 1$  if age 20 - 22  
 $T_4 = 1$  if age 22 - 27  
 $T_5 = 1$  if age 27-32  
 $T_6 = 1$  if age is greater than 32 years.

\* The upper limit of each group is not inclusive -- i.e., 16-20 means 16 to 19.999 ....

The above dummy variables are exhaustive and mutually exclusive -- i.e., one must occur, and no other will occur simultaneously. The division of the ordinary variables into classes may be accomplished arbitrarily (e.g., 2 year steps) or intuitively from prior experience (as in the above example). A caution is in order. The predictor screening technique will not generate finer divisions, so in starting out one should not combine possibly significant groups. The screening procedure will safely accomplish this task and account for most of the predictive information.

Let the initial set of tentative dummy predictors under consideration be designated as  $T_1, T_2, \dots, T_Q$ . The screening involves the computation of variance-ratio statistics  $F$ , where an individual  $F$  tests the significance of an additional predictor. If  $R_k$  denotes the multiple correlation coefficient computed from  $k$  predictors, and d.f. stands for the estimated number of degrees of freedom left at stage  $k$ , the variance ratio  $F$  used to test the  $k$ -th selection is given by the equation

$$F = \frac{R_k^2 - R_{k-1}^2}{1 - R_k^2} \cdot (\text{d. f.}) \quad (5-12)$$

Assuming independent observations, the value of d.f. for the  $k$ -th selection is  $N - k - 1$ , where  $N$  is the sample size. See Miller (1964) for a proper level of significance for  $F$ .

Screening accounts for all  $G$  predictand states simultaneously and is done as follows. Compute a value of  $F$  for each tentative predictor  $T_q$  ( $q = 1, 2, \dots, Q$ ) in relation to each dummy predictand  $D_g$  ( $g = 1, 2, \dots, G$ ). At the first stage of predictor selection, there will be  $G \times Q$  values of  $F$  (since  $G$  values will be obtained for each  $T$ ). Denote by  $x^1$  the predictor that yields the largest single value of  $F$  out of all of these  $G \times Q$  values. This predictor  $x^1$  is called the first predictor.

The screening process is now repeated to select a second predictor. For each  $D_g$ , trial multiple correlations using two predictors are computed. The two predictors on any trial are  $x^1$  and one of the remaining  $T$ 's. There will be  $G(Q - 1)$  such multiple correlations with the same number of  $F$ -values. The trial predictor yielding the largest value of  $F$  among these  $G(Q - 1)$  values is selected as the second predictor and is denoted by  $x^2$ .

The screening is continued to select third, fourth, and further predictors until  $S$  predictors  $x^1, x^2, \dots, x^S$  have been chosen. As each predictor is selected, a statistical test comparing the highest computed  $F$ -value with a certain critical value of  $F$  is employed to decide whether the proposed, selected predictor appears to be useful. The termination point  $S$  is established by the fact that  $x^S$  passes this test, but the next candidate (which, if successful, would be called  $x^{S+1}$ ) fails it.

The process just described is called forward screening to distinguish it from a different, but related, selection process, called backward screening. In backward screening, a definite set of  $B$  ( $B = Q$ ) trial predictors is chosen to begin with, and a regression formula based on all  $B$  predictors is determined. The least important predictor is then identified by calculating the increase in mean square error due to the omission of each predictor, in turn, when the other  $B - 1$



predictors are retained. If the least important predictor is judged non-significant, it is eliminated. The tests are applied again to the remaining set of  $B - 1$  predictors, and the deletion process is continued in a stepwise fashion, analogous to that used in forward screening. Because forward screening can cope with a much larger set of tentative predictors than can backward screening, it was chosen as the method to be utilized in REEP.

The ordinary formulas for estimating the sampling variability of regression constants do not hold when the predictors are required to meet preliminary tests of significance, as they are in selective screening. The most important single question to answer is not that of the sampling behavior of separate coefficients, but rather that of the sampling behavior of the estimated regression function as a whole. In REEP, the latter question is attacked by reserving an independent sample for verification.

By a randomization technique, the initial sample is divided into two parts. One part, usually the larger, is called the developmental sample and is used for all processes involving setting up the problem -- objective dummyming, predictor screening, fitting of constants. The other part, called the verification sample, is used solely to obtain estimates of predictive accuracy when the regression formulas are applied to independent data. The program can accept a developmental sample size of about 10,000 and, if desired, an even larger sample size for verification.

Predictive performance is measured by the correspondence between  $\hat{D}_g$  and  $D_g$  in the verification sample. ( $\hat{D}_g$  reduces to  $D_g$  if no adjustment is required because probabilities cannot be less than zero or greater than unity).

An overall measure of correspondence between  $\hat{D}_g$  and  $D_g$  is given by the mean-square error, as defined by the Brier P-Score. For a single probability forecast of  $G$  states, the P-Score is defined as

$$P\text{-Score} = \sum_{g=1}^G (\hat{D}_g - D_g)^2 \quad (5-13)$$

A P-Score of 0.0 indicates a perfect forecast; the poorest score is 2.0, which results when for some value  $g$ ,  $D_g = 1$ , whereas in fact there exists some other value  $g'$  such that  $D_{g'} = 1$ . In comparing two forecasts of the same events, the lower P-Score indicates the better forecast. For a series of  $N$  probability forecasts of  $G$  states, the P-Score is defined as follows:

$$P\text{-Score} = \frac{1}{N} \sum_{i=1}^N \sum_{g=1}^G (\hat{D}_{gi} - D_{gi})^2 \quad (5-14)$$

Bryan and Singer used two types of variables in this application of REEP -- univariate and bivariate.\* In one model, Model A, only univariates were allowed to be selected for predicting reenlistment action, while in their second model, Model B, both univariates and bivariate were considered. Because of the completeness of their work, we can evaluate how much, if any, improvement in prediction is obtained by including bivariate as well as univariate.

Of 61 dummy variables under consideration as possible predictors in Model A, seven were selected as significant by the screening procedure. Therefore, the REEP regression function for estimating reenlistment rate is of the form:

$$\hat{D}_1 = B_0 + B_1 x^1 + \dots + B_7 x^7 \quad (5-15)$$

The selected predictors (designated by  $x$ 's) and values of the regression coefficients ( $B$ 's) are given in Table 5-1. The notation  $x^1$  represents the most significant predictor,  $x^2$  the second most significant predictor, or more precisely the most significant adjunct to  $x^1$ , and so on in order of superscript, so that  $x^7$  represents the seventh most significant.

\* univariate are used here to mean dummy variable predictors while bivariate are joint dummy predictors.

TABLE 5-1  
TERMS IN  $\hat{D}_1$  FOR MODEL A

Predictor Symbol		Additive Constant	$x^1$	$x^2$	$x^3$	$x^4$	$x^5$	$x^6$	$x^7$
Regression Coefficients	Symbol	$B_0$	$B_1$	$B_2$	$B_3$	$B_4$	$B_5$	$B_6$	$B_7$
	Value	.272	-.063	.292	.075	.059	.281	-.047	.082

To use the data in Table 5-1 for deriving a reenlistment rate for any given individual is a relatively simple matter. The coefficients of the characteristics that pertain to the individual are added to the baseline, the "additive constant". For example, if none of the seven selected categories pertain to an individual, then his predicted reenlistment rate is simply .272, which is a little above average. If categories 1, 2, and 4 pertain, then his predicted probability is .560 (.272 - .063 + .292 + .059), which is quite high.

The Brier P-Score for the developmental sample (N=6372) was .3703; that for the verification sample (N=703) was .3707. Hence, overall predictive performance, as measured by the Brier P-Score, was very nearly the same on the verification sample as on the developmental sample, thus lending credence to the method. If the population itself were to undergo basic changes in the relationships among the variables, the present regression function should not be expected to apply.

In Model B, both univariate and bivariate predictors were made available for selection. The results using Model B parallel those of Model A.

Of 214 dummy variables (61 univariates and 153 bivariates) considered as possible predictors in Model B, seven were selected as significant by the screening procedure. The REEP regression function for estimating reenlistment rate thus reduced to the same form as in Model A.

$$\hat{D}_1 = C_0 + C_1 y^1 + \dots + C_7 y^7 \quad (5-16)$$

Both the selected predictors and the regression coefficients differed from those derived in Model A. The selected predictors (designated by y's) and values of the coefficients (C's) are shown in Table 2. As in Model A, the subscripts indicate the rank order of selection.

TABLE 5-2  
TERMS IN  $\hat{D}_1$  FOR MODEL B

Predictor Symbol		Additive Constant	$y_1$	$y_2$	$y_3$	$y_4$	$y_5$	$y_6$	$y_7$
Regression Coefficients	Symbol	$C_0$	$C_1$	$C_2$	$C_3$	$C_4$	$C_5$	$C_6$	$C_7$
	Value	.257	-.068	.393	-.053	.096	.338	.056	.071

The Brier P-Score for the developmental sample of Model B (N = 6372) was 0.3687; that for the verification sample (N = 703) was 0.3694. The indicated superiority of Model B over Model A, shown by the slightly (but statistically significantly) lower P-Score of the former in the developmental sample, was verified in the verification sample.

Bryan and Singer showed that both models yield valid estimates of reenlistment probability, but that Model B had greater capacity for sorting out departures from average. On comparing the predictors selected under the two models, it was found

that six of the seven selected predictors were closely related.

A study, such as that performed by Bryan and Singer, will remain valid as long as there are no significant changes in either the predictors or their relative weight. Consequently, the use of such a scheme should employ frequent verification as an indicator of the need to update the model.



## CHAPTER 6

### CANONICAL CORRELATION APPLICATIONS

by Lt Jeanette M. Heumann

#### INTRODUCTION

The key to marketing success lies in matching a buyer with a product which will fill his or her needs. A buyer study provides a concrete means of diagnosing the buyer-product relationship. A buyer study will statistically relate buyer characteristics to product characteristics so that for a given buyer, a probability distribution may be predicted as to the size and type of that buyer's purchase. A study such as this, which examines buyer characteristics versus those of a particular product, is an important management information tool. An in depth statistical analysis of this relationship can provide a basis for company goal setting and planning. It will point out a company's strengths and weaknesses in the area of their marketing operations. A buyer study should examine several areas. It should tell who your buyers are and what products they are likely to buy. Buyers may be identified by variables such as buyer sex, age, income, marital status, number of dependents, and occupation. A buyer study also allows a company to evaluate its product performance as compared to peer companies' product performance. In addition, a buyer study will identify a company's principal markets. Markets are defined as groups of individuals who possess homogeneous characteristics such as marital status, occupation, income, or age. Finally, a company buyer study will clearly indicate, by means of a statistical study, what is sold to a company's principal markets.

Buyer studies have long been used by insurance companies to evaluate new markets or territories, in selective advertising, in persistency evaluation, in quota setting, in appraising new product potential, and in assisting new agents. The means by which a set of buyer characteristics or variables are related to a set of product variables is through the use of statistical procedures such as canonical correlation. This statistical procedure was initially developed by Hotelling (1935).

Insurance companies can collect "families" of information about buyers (sex, income, age,... etc.). Canonical correlation methods treat these "families" of information as separate entities and yet allow the probability of occurrence of one type of family information to be calculated from another family of information.

This paper will examine work done by Dr. R. G. Miller for the Life Insurance Marketing and Research Association (LIMRA) which discusses the buyer-product relationship as applied to the insurance industry. This work relates buyer variables such as:

- Marital status
- Age
- Resident state
- Income
- Occupation
- Age and income combined
- Age, income, and sex combined
- Income and occupation combined

to product variables such as:

- Mode of payment
- Type policy
- Amount of policy
- Policy premiums
- Type of policy combined with mode of payment

Approximately twenty thousand United States ordinary policies sold to adults in 1970 were used in analyzing the probability of a particular individual buying one of a number of policies which differed in size and type.

#### Terms and Mathematical Symbols

The following is a list of terms and symbols which will be used extensively in this paper and with which the reader should be familiar.

1. M - Underlining signifies a matrix
2.  $x_p$  - Variable describing a product characteristic
3.  $y_q$  - Variable describing a buyer characteristic
4.  $a_{tp}$  - coefficient of  $x$
5.  $b_{tq}$  - coefficient of  $y$
6. a and b - two sets of weights which maximize the correlation between the derived canonical variates

where  $\underline{a} = \begin{matrix} & \begin{matrix} 1 & 2 & \dots & p \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ \vdots \\ t \end{matrix} & \begin{bmatrix} a_{11} & & & \\ & a_{22} & & \\ & & \ddots & \\ & & & a_{tp} \end{bmatrix} \end{matrix}$  and  $\underline{b} = \begin{matrix} & \begin{matrix} 1 & 2 & \dots & q \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ \vdots \\ t \end{matrix} & \begin{bmatrix} b_{11} & & & \\ & b_{22} & & \\ & & \ddots & \\ & & & b_{tq} \end{bmatrix} \end{matrix}$

7.  $V_1, V_2, \dots, V_t$  - Linear functions, where each  $V$  represents a linear combination of  $x$  variables
8.  $W_1, W_2, \dots, W_t$  - Linear functions, where each  $W$  represents a linear combination of  $y$  variables
9.  $\underline{R}$  - Correlation matrix where  $\underline{R} = \begin{bmatrix} \underline{R}_{11} & \underline{R}_{12} \\ - & - \\ \underline{R}_{21} & \underline{R}_{22} \end{bmatrix}$
10.  $\underline{R}_{11}$  - The intercorrelation among the  $x$ 's
11.  $\underline{R}_{22}$  - The intercorrelation among  $y$ 's
12.  $\underline{R}_{12}$  - The intercorrelation of  $x$ 's and  $y$ 's
13.  $\underline{R}_{21}$  - The transpose of  $\underline{R}_{12}$
14.  $\lambda$  - Lagrange multiplier
15.  $\Lambda$  - Wilk's criterion, a likelihood-ratio criterion
16. SSCP Matrix (S) - The sums of squares and cross products matrix; see Tatsuoaka (1971)
17. WLLP - A whole life - limited pay insurance policy
18. WLCP - A whole life-continuous pay insurance policy
19. MODL - A modified life insurance policy
20. ENDR - An endowment and retirement insurance policy
21. LEVT - Level term insurance policy, DECT - Decreasing term insurance policy
22. COMB - A combination policy
23. *A posteriori* probabilities - Conditional probabilities over the states of nature, or predictand groups for given predictor values.
24. *A priori* probabilities - Unconditional probabilities over predictand groups of the states of nature
25. Mahalanobis  $D^2$  - A statistic utilized for testing the significance of  $P$  variates to discriminate among two groups which have equal or unequal dispersion but, different means.
26. Unit matrix - A matrix which is symmetric and has diagonal elements equal to unity while all other elements equal zero

### Technical Details

As was mentioned earlier, the particular insurance case study being examined used approximately 20,000 ordinary policies sold in the United States in 1970. The study excluded from analysis the following types of policies:

Policies on lives of residents of U.S. territories,  
Canada, and other foreign countries,  
Acquired reinsurance cases,  
Individual credit insurance,  
Group insurance,  
Annuities without insurance,  
Term conversions,  
Single or double premium policies.

By examining the nonexcluded policies, it was possible to determine two sets of variables. The first set of variables (x's) described product characteristics and made up the set of predictand variables. The second set of variables described buyer characteristics and made up the set of predictor variables. The interrelations between these two sets of measurements can be studied by canonical correlation models. This statistical method provides the maximum correlation between the linear functions of the two sets of variables. Each pair of these linear functions is determined so that the correlation between the new pair of canonical variates is maximized where the maximization of correlation is subject to the limitation that they are independent of linear combinations derived previously.

The general nature of canonical correlation may be best explained using algebraic means. First, consider two simultaneous sets of  $t$  equations that contain  $p$  predictand and  $q$  predictor variables. From the two sets of equations as follows:

"Left Side" (Predictand Equations)

$$\begin{aligned}V_1 &= a_{11}x_1 + a_{12}x_2 + \dots + a_{1p}x_p \\V_2 &= a_{21}x_1 + a_{22}x_2 + \dots + a_{2p}x_p \\&\vdots \\V_t &= a_{t1}x_1 + a_{t2}x_2 + \dots + a_{tp}x_p\end{aligned}$$

"Right Side" (Predictor Equations)

$$\begin{aligned}W_1 &= b_{11}y_1 + b_{12}y_2 + \dots + b_{1q}y_q \\W_2 &= b_{21}y_1 + b_{22}y_2 + \dots + b_{2q}y_q \\&\vdots \\W_t &= b_{t1}y_1 + b_{t2}y_2 + \dots + b_{tq}y_q\end{aligned}$$

(where  $t = \min(p, q)$ )

We can then determine the sets of weights of  $a$  and  $b$  so that the correlation between  $V_1$  and  $W_1$  is higher than that of any pair of linear functions of the  $x$ 's and the  $y$ 's. Likewise  $V_2$  and  $W_2$  are correlated higher than any pair of simultaneous equations other than  $V_1$  and  $W_1$ . This hierarchy of correlation is maintained all the way down to the  $V_t$  and  $W_t$  equations. For this relationship to exist, the weights of  $a$  and  $b$  must be determined so as to maximize the relationship between the derived canonical variates  $\bar{V}$  and  $\bar{W}$ . It should be noted that the special case where  $q > 1$  and  $p = 1$  is a case of multiple regression and not one of canonical correlation. Canonical correlation requires that there be both multiple predictors and multiple predictands involved. The size of  $p$  or  $q$  determines the number of linear combinations that can be formed. If  $q$  is smaller than  $p$  then, there will be  $q$  linear combinations formed. Likewise if  $p$  is smaller than  $q$ , then  $p$  linear combinations will be formed. Each pair of the canonical variates  $V$  and  $W$  will have maximum correlation, taking into account the restriction that each canonical variate ( $V_i$  or  $W_i$ ) is orthogonal to all other canonical variates on its side of the equation.

In geometric terms we can consider canonical correlation as a measure of the extent to which individuals can occupy the relative positions in the  $p$  dimensional space as they do in  $q$  dimensional space. Considering buyer-predictor variables versus the insurance product-predictand variables might appear to possess little or no similarity when the variables are compared scale



for scale. However, canonical correlation methods readily show the system of correlation which underlies the two sets of variables. The first step in the analysis of the canonical correlation between the set of buyer variables and the insurance product variables is the solving of a and b. This requires the correlation matrix R. Where R equals:

$$\begin{bmatrix} \underline{R}_{11} & | & \underline{R}_{12} \\ \hline \underline{R}_{21} & | & \underline{R}_{22} \end{bmatrix}$$

As can be readily seen from the above matrix, R is itself divided into four submatrices. The four submatrices are obtained in the following manner:

R is the correlation between the x's

$$\underline{R}_{11} = \begin{matrix} & \begin{matrix} x_1 & x_2 & x_3 & \dots & x_p \end{matrix} \\ \begin{matrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_p \end{matrix} & \begin{bmatrix} 1 & r_{x_1 x_2} & r_{x_1 x_3} \dots & r_{x_1 x_p} \\ & 1 & r_{x_2 x_3} \dots & r_{x_2 x_p} \\ & & 1 & \dots & r_{x_3 x_p} \\ & & & \ddots & \vdots \\ & & & & 1 \end{bmatrix} \end{matrix}$$

Note: This matrix is represented as an upper right triangle matrix due to the fact that the x-x correlations composing the lower left triangle of the matrix are simply mirror images of the upper right triangle correlations.

R<sub>22</sub> is the correlation among the y's

$$\underline{R}_{22} = \begin{matrix} & \begin{matrix} y_1 & y_2 & y_3 & \dots & y_q \end{matrix} \\ \begin{matrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_q \end{matrix} & \begin{bmatrix} 1 & r_{y_1 y_2} & r_{y_1 y_3} \dots & r_{y_1 y_q} \\ & 1 & r_{y_2 y_3} \dots & r_{y_2 y_q} \\ & & 1 & \dots & r_{y_3 y_q} \\ & & & \ddots & \vdots \\ & & & & 1 \end{bmatrix} \end{matrix}$$

Again, the R<sub>22</sub> matrix (like the R<sub>11</sub> matrix) may be displayed as an upper right triangle matrix.

R<sub>21</sub> is the correlation between the x's and the y's

$$\underline{R}_{21} = \begin{matrix} & \begin{matrix} x_1 & x_2 & x_3 & \dots & x_p \end{matrix} \\ \begin{matrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_q \end{matrix} & \begin{bmatrix} r_{y_1 x_1} & r_{y_1 x_2} & r_{y_1 x_3} \dots & r_{y_1 x_p} \\ r_{y_2 x_1} & r_{y_2 x_2} & r_{y_2 x_3} \dots & r_{y_2 x_p} \\ r_{y_3 x_1} & r_{y_3 x_2} & r_{y_3 x_3} \dots & r_{y_3 x_p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{y_q x_1} & r_{y_q x_2} & r_{y_q x_3} \dots & r_{y_q x_p} \end{bmatrix} \end{matrix}$$

And finally,  $\underline{R}_{12}$  which is simply the transpose of  $\underline{R}_{21}$ . That is

$$\underline{R}_{12} = \begin{matrix} & \begin{matrix} y_1 & y_2 & y_3 & \dots & y_q \end{matrix} \\ \begin{matrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_p \end{matrix} & \begin{bmatrix} r_{x_1,y_1} & r_{x_1,y_2} & r_{x_1,y_3} & \dots & r_{x_1,y_q} \\ r_{x_2,y_1} & r_{x_2,y_2} & r_{x_2,y_3} & \dots & r_{x_2,y_q} \\ r_{x_3,y_1} & r_{x_3,y_2} & r_{x_3,y_3} & \dots & r_{x_3,y_q} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{x_p,y_1} & r_{x_p,y_2} & r_{x_p,y_3} & \dots & r_{x_p,y_q} \end{bmatrix} \end{matrix}$$

The partitioned portions of  $\underline{R}$  may then be substituted into the canonical equation

$$(\underline{R}_{11}^{-1} \quad \underline{R}_{12} \quad \underline{R}_{22}^{-1} \quad \underline{R}_{21} - \lambda_i^2 \underline{I}) \underline{a}_i = 0.$$

NOTE: the above equation may also be written as

$$(\underline{R}_{22}^{-1} \quad \underline{R}_{21} \quad \underline{R}_{11}^{-1} \quad \underline{R}_{12} - \lambda_i^2 \underline{I}) \underline{b}_i = 0.$$

The equation for  $\underline{a}_i$  may be derived through the following procedure:

1. Let  $c_i = \lambda_i$  (where  $i = 1, 2, \dots, \min(p, q)$ ).
2. Since we are looking for linear functions that have maximum correlation and since the correlation of a multiple of  $\underline{V}$  and a multiple of  $\underline{W}$  is the same as the correlation of  $\underline{V}$  and  $\underline{W}$  we can make an arbitrary normalization of  $\underline{a}$  and  $\underline{b}$ .

This is done such that:

$$\begin{aligned} 1 &= \underline{a}' \underline{V} \underline{a} = \underline{a}' \underline{R}_{11} \underline{a} \\ 1 &= \underline{b}' \underline{W} \underline{b} = \underline{b}' \underline{R}_{22} \underline{b} \end{aligned}$$

3.  $\underline{a}' \underline{R}_{11} \underline{a} = \underline{b}' \underline{R}_{22} \underline{b} = 1$ .
4.  $\underline{V} \underline{W} = \underline{a}' \underline{R}_{12} \underline{b}$ .
5. Since the algebraic problem is to maximize (4) subject to (3), let

$$F(\underline{a}, \underline{b}) = (\underline{a}' \underline{R}_{12} \underline{b} - \lambda/2)(\underline{a}' \underline{R}_{11} \underline{a} - 1) - \mu/2(\underline{b}' \underline{R}_{22} \underline{b} - 1)$$

where  $\lambda$  and  $\mu$  are Lagrange multipliers and where the factors  $1/2$  are introduced for numerical convenience.

6. Then if we differentiate  $F$  with respect to  $\underline{a}$  and  $\underline{b}$  and set the vectors of the derivatives equal to zero, we have:

$$\frac{\partial F}{\partial \underline{a}} = \underline{R}_{12} \underline{b} - \lambda \underline{R}_{11} \underline{a} = 0$$

$$\frac{\partial F}{\partial \underline{b}} = \underline{R}_{21} \underline{a} - \mu \underline{R}_{22} \underline{b} = 0$$

multiply  $\frac{\partial F}{\partial \underline{a}}$  by  $\underline{a}'$ , and  $\frac{\partial F}{\partial \underline{b}}$  by  $\underline{b}'$

$$\text{so that, } \underline{a}' \underline{R}_{12} \underline{b} - \lambda \underline{a}' \underline{R}_{11} \underline{a} = 0$$

$$\underline{b}' \underline{R}_{21} \underline{a} - \mu \underline{b}' \underline{R}_{22} \underline{b} = 0.$$

7. Since we know that  $\underline{a}' \underline{R}_{11} \underline{a} = 1$ , and  $\underline{b}' \underline{R}_{22} \underline{b} = 1$ , we can easily see that

$$\lambda = \mu = \underline{a}' \underline{R}_{12} \underline{b}.$$

8. Thus we can write: a)  $-\lambda_i R_{i-11} \underline{a}_i + R_{i-12} \underline{b}_i = 0$

$$b) \underline{R}_{i-21} \underline{a}_i - \lambda_i \underline{R}_{i-22} \underline{b}_i = 0$$

9. We can then derive a single matrix equation for  $\underline{a}_i$  or  $\underline{b}_i$  if we multiply 8a by  $\lambda_i$  and 8b by  $R_{i-22}^{-1}$ .

$$a) \lambda_i \underline{R}_{i-12} \underline{b}_i = \lambda_i^2 \underline{R}_{i-11} \underline{a}_i$$

$$b) \underline{R}_{i-22}^{-1} \underline{R}_{i-21} \underline{a}_i = \lambda_i \underline{b}_i$$

10. If we then substitute from 9a into 9b...

$$a) \underline{R}_{i-12} \underline{R}_{i-22}^{-1} \underline{R}_{i-21} \underline{a}_i - \lambda_i^2 \underline{R}_{i-11} \underline{a}_i = 0$$

$$b) (\underline{R}_{i-12} \underline{R}_{i-22}^{-1} \underline{R}_{i-21} - \lambda_i^2 \underline{R}_{i-11}) \underline{a}_i = 0$$

$$c) (\underline{R}_{i-11} \underline{R}_{i-12} \underline{R}_{i-22}^{-1} \underline{R}_{i-21} - \lambda_i^2 \underline{I}) \underline{a}_i = 0$$

Where  $|\underline{R}_{i-11} \underline{R}_{i-12} \underline{R}_{i-22}^{-1} \underline{R}_{i-21} - \lambda_i^2 \underline{I}| = 0$  and  $\underline{a}_1, \dots, \underline{a}_p$  satisfy equation 10a for  $\lambda_i^2 = \lambda_1^2, \dots, \lambda_p^2$  respectively. The similar equations for  $\underline{b}_1, \dots, \underline{b}_q$  occur when  $\lambda_i^2 = \lambda_1^2, \dots, \lambda_q^2$  are substituted with

$$(\underline{R}_{i-22}^{-1} \underline{R}_{i-21} \underline{R}_{i-11} \underline{R}_{i-12} - \lambda_i^2 \underline{I}) \underline{b}_i = 0$$

Note: the vector  $\underline{b}_i$  may also be obtained from the equation

$$\underline{b}_i = (\underline{R}_{i-22}^{-1} \underline{R}_{i-21} \underline{a}_i) / \lambda_i$$

The vectors  $\underline{a}_i$  and  $\underline{b}_i$  are then applied to standard score vectors to obtain the canonical variates V and W. The canonical correlation ( $R_c$ ) between the  $i$ th pair of new composites is equal to  $\lambda_i$ . The largest  $\lambda_i^2$  is the square of the maximum possible correlation between the linear combinations of the two sets of measurements ( $R_c^2 \max = \lambda_i^2$ ). Also, if it is desired to find the coefficients of the observed deviation scores, they can be obtained by dividing the elements  $\underline{a}_i$  and  $\underline{b}_i$  by the standard deviation of the corresponding variables.

An alternate but similar procedure for finding the canonical correlation between two sets of variables, in this case buyer variables and product variables, is discussed by Tatsuoka in his book, Multivariate Analysis: Techniques for Educational and Psychological Research. Tatsuoka considers two sets of variables which each construct a linear combination:

$$V = a_1 x_1 + a_2 x_2 \dots a_p x_p$$

$$W = b_1 y_1 + b_2 y_2 \dots b_q y_q$$

From these linear functions, we must determine the two sets of coefficients  $a' = a_1, a_2, \dots, a_p$  and  $b' = b_1, b_2, \dots, b_q$  so as to maximize correlation between the two linear combinations.<sup>2</sup> In order to do this we must express the correlation between V and W as a function of  $\underline{a}$  and  $\underline{b}$ .

$$r_{vw} = EVW / [(EV^2) (EW^2)]^{1/2}$$

We may express the quantities  $V^2$  and  $W^2$  as quadratic forms in the following manner:

$$EV^2 = \underline{a}' \underline{S}_{xx} \underline{a}$$

$$EW^2 = \underline{b}' \underline{S}_{yy} \underline{b}$$

The quantities  $\underline{S}_{xx}$  and  $\underline{S}_{yy}$  represent the Sums of Squares and Cross Products Matrices (SSCP). This type matrix is mentioned in the Terms section of this paper and explained fully by Tatsuoka (1971).

We can also show that:  $EVW = \underline{a}' \underline{S}_{xy} \underline{b}$ . Where in this case  $\underline{S}_{xy}$  represents the  $p \times q$  matrix of the sums-of-products between the x variables and the y variables. Using the above equalities, the formula for  $r_{vw}$  may be written in the following manner:

$$r_{vw} = \underline{a}' \underline{S}_{xy} \underline{b} / [(\underline{a}' \underline{S}_{xx} \underline{a}) (\underline{b}' \underline{S}_{yy} \underline{b})]^{1/2}$$

The maximizing weights of  $\underline{a}$  and  $\underline{b}$  are determined only up to proportionality constants. This is done because, if  $\eta$  and  $\upsilon$  are two arbitrary constants of the same sign, the value of the correlation



between V and W obtained by using the elements of  $\underline{a}$  and  $\underline{b}$  as combining weights is seen to be equal to the value which would result from the use of  $a_i$  and  $b_i$  as the weights.

The proportionality constants can be chosen so that:

$$\underline{a}' \underline{S}_{xx} \underline{a} = 1$$

$$\underline{b}' \underline{S}_{yy} \underline{b} = 1$$

This results in the denominator of  $\underline{a}' \underline{S}_{xy} \underline{b} / [(\underline{a}' \underline{S}_{xx} \underline{a})(\underline{b}' \underline{S}_{yy} \underline{b})]^{1/2}$  being equal to unity. If Lagrange multipliers  $\lambda/2$  and  $\mu/2$  are introduced at this time (the factors  $1/2$  are introduced merely for numerical convenience), the the function which is to be maximized is

$$F(\underline{a}, \underline{b}) = \underline{a}' \underline{S}_{xy} \underline{b} - (\lambda/2) (\underline{a}' \underline{S}_{xx} \underline{a} - 1) - (\mu/2) (\underline{b}' \underline{S}_{yy} \underline{b} - 1).$$

The next step involves taking the symbolic partial derivatives of  $F(\underline{a}, \underline{b})$  with respect to  $\underline{a}$  and  $\underline{b}$ . The two resulting equations are then each set equal to the null vector. This gives the following equations:

$$\frac{\partial F}{\partial \underline{a}} = \underline{S}_{xy} \underline{b} - \lambda \underline{S}_{xx} \underline{a} = 0$$

$$\frac{\partial F}{\partial \underline{b}} = \underline{a}' \underline{S}_{xy} - \mu \underline{b}' \underline{S}_{yy} = 0$$

which are the equations that must be satisfied by  $\underline{a}$  and  $\underline{b}$  in order to maximize the correlation coefficient  $r_{vw}$ . The above two equations constitute sufficient conditions for the desired maximization.

The next step is to premultiply the members of  $\partial F / \partial \underline{a}$  by  $\underline{a}'$  and to postmultiply the members of  $\partial F / \partial \underline{b}$  by  $\underline{b}$  as follows:

$$\underline{a}' \underline{S}_{xy} \underline{b} - \lambda (\underline{a}' \underline{S}_{xx} \underline{a}) = 0$$

$$\underline{a}' \underline{S}_{xy} \underline{b} - \mu (\underline{b}' \underline{S}_{yy} \underline{b}) = 0$$

From the preceding two equations the relationship of  $\underline{a}' \underline{S}_{xy} \underline{b} = \lambda (\underline{a}' \underline{S}_{xx} \underline{a}) = \mu (\underline{b}' \underline{S}_{yy} \underline{b})$  can easily be seen. This relationship reduces to the form

$$\underline{a}' \underline{S}_{xy} \underline{b} = \lambda = \mu \quad \text{by recalling that } \underline{a}' \underline{S}_{xx} \underline{a} = \underline{b}' \underline{S}_{yy} \underline{b} = 1$$

This clearly shows that both eigenvalues  $\lambda$  and  $\mu$  are equal to the maximum value that can be achieved by the correlation coefficient  $r_{vw}$ . Since  $\lambda = \mu$  we may replace  $\lambda$  by  $\mu$  so that:

$$\underline{S}_{xy} \underline{b} - \mu \underline{S}_{xx} \underline{a} = 0$$

$$\underline{S}_{xy} \underline{b} = \mu \underline{S}_{xx} \underline{a}$$

$$\text{and } \underline{S}_{yx} \underline{a} - \mu \underline{S}_{yy} \underline{b} = 0$$

$$\underline{S}_{yx} \underline{a} = \mu \underline{S}_{yy} \underline{b}$$

If we then assume  $S_{yy}$  to be nonsingular, we may express  $\underline{b}$  in terms of  $\underline{a}$  as follows:

$$\underline{b} = (1/\mu) S_{yy}^{-1} S_{yx} \underline{a}.$$

This expression may then be substituted for  $\underline{b}$ :

$$S_{xy} [(1/\mu) S_{yy}^{-1} S_{yx} \underline{a}] = \mu S_{xx} \underline{a}.$$

If we then premultiply both members of the above equation by  $\mu S_{xx}^{-1}$  we obtain the following expressions:

$$\begin{aligned} S_{xx}^{-1} S_{xy} S_{yy}^{-1} S_{yx} \underline{a} &= \mu^2 \underline{I} \underline{a} \\ S_{xx}^{-1} S_{xy} S_{yy}^{-1} S_{yx} \underline{a} - \mu^2 \underline{I} \underline{a} &= 0 \end{aligned}$$

and in final form,  $(S_{xx}^{-1} S_{xy} S_{yy}^{-1} S_{yx} - \mu^2 \underline{I}) \underline{a} = 0.$

From this equation it can be seen that the largest eigenvalue  $\mu^2$ , of the quadruple matrix product  $S_{xx}^{-1} S_{xy} S_{yy}^{-1} S_{yx}$  gives the square of the maximum correlation coefficient. It can also be seen that the elements of the associated eigenvector  $\underline{a}_1$  provide the weights by which the x set of variables need to be combined linearly to achieve this maximum correlation. The other combining weight vector  $\underline{b}_1$  may be easily obtained by substituting  $\underline{a}_1$  and  $\mu_1$  in the equation

$$\underline{b}_1 = (1/\mu_1) S_{yy}^{-1} S_{yx} \underline{a}_1.$$

It is obvious then that other eigenvalues and vectors besides  $\mu_1^2$  and  $\underline{a}_1$  may be obtained using the general equations just developed. These equations may be utilized to determine weights  $\underline{a}$  and  $\underline{b}$  when two sets of linear functions  $V_1, V_2, \dots, V_t$  and  $W_1, W_2, \dots, W_t$  are present. We start out by finding two sets of combining weights that will maximize the specified correlation coefficient for the resulting pair of linear combinations. The elements of the vector which are associated with the largest eigenvalue of a certain matrix make up the weights which lead to the desired absolute maximum. This also holds true of the elements of the vectors associated with the second, third, and final eigenvalues in descending order of magnitude in the sense that the pair of linear combinations formed by the elements of the second vector has the largest value of the relevant criterion among those that are not correlated with the first pair of linear combinations and so forth. For any case of canonical correlation analysis, the linear combinations occur in two sequences, one sequence for each of the two sets of variables. There is no correlation within each sequence and between unmatched pairs of linear combinations across the two sequences. Therefore, not only is  $V_1$  uncorrelated with  $V_2, V_3, \dots$  or  $V_t$ , it is also uncorrelated with  $W_2, W_3, \dots$  or  $W_t$ . It should be clear then that the only nonzero correlations occur between the corresponding members of the paired linear combinations such as  $V_1$  and  $W_1$ , or  $V_2$  and  $W_2$ . Canonical variates is the term used to denote pairs of linear orthogonal combinations. The number of canonical variate pairs will equal the number of variables in the smaller variable set, that is the smaller of the two numbers  $p$  or  $q$ . This is due to the fact that the rank of  $S_{xx}^{-1} S_{xy} S_{yy}^{-1} S_{yx}$ , the quadruple product matrix, whose eigenvectors and eigenvalues determine the canonical variates, is equal to  $q$  or  $p$  whichever is the smaller.

An application of Bayes' Theorem is utilized to evaluate *a posteriori* probabilities. Specifically, this means that all of the possible "Left Side" combinations ( $L$ ) will be enumerated for which it is desired to obtain probabilities. There will be  $L \leq 2^p$  "Left Side" combinations. This may be performed for the set of "Right Side" conditions of  $\underline{w} = w_1, w_2, w_3, \dots, w_t$  by utilizing the following equation:

$$P(L|\underline{w}) = \frac{f(\underline{w}|L) g_L}{\sum_{L=1}^L f(\underline{w}|L) g_L} \quad (\text{where } L = 1, 2, 3, \dots, L)$$

and  $g_\ell$  is the *a priori* probability of  $\ell$ . In this equation  $f(w|\ell)$  is assumed to be multivariate normal for all  $\ell$  combinations with equal covariance matrix  $\underline{\Sigma}$ , or

$$f(w|\ell) = \frac{|\underline{\Sigma}|^{-\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}} \exp \left[ -\frac{1}{2} (\underline{w} - \underline{\hat{w}}_\ell)' \underline{\Sigma}^{-1} (\underline{w} - \underline{\hat{w}}_\ell) \right]$$

where  $\ell = 1, 2, 3, \dots, L$  and  $\underline{\hat{w}}_\ell$ , the vector of values that are expected to be predicted from the "Left Side" for the combination. The diagonal matrix  $\underline{\Sigma}$  has elements whose values represent the error variances between the pairs of linear combinations  $V_1$  and  $W_1$ ,  $V_2$  and  $W_2$ ,  $V_3$  and  $W_3, \dots, V_t$  and  $W_t$ . If on the other hand, we desire to find the probabilities for a set of "Right Side"  $R$  combinations, we can do so in the following manner: There will be  $R \leq 2^p$  "Right Side" combinations. To perform this for the set of "Left Side" conditions of  $v = v_1, v_2, \dots, v_t$  by utilizing the following equation

$$p(k|v) = \frac{f(v|k)g_k}{\sum_{k=1}^R f(v|k)g_k}$$

where  $k=1,2,3,\dots,R$  and  $g_k$  is the *a priori* probability of  $k$ , and  $f(v|k)$  is again assumed to be multivariate normal for all  $R$  combinations with equal covariance matrix  $\underline{\Sigma}$  or

$$f(v|k) = \frac{|\underline{\Sigma}|^{-\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}} \exp \left[ -\frac{1}{2} (\underline{v} - \underline{\hat{v}}_k)' \underline{\Sigma}^{-1} (\underline{v} - \underline{\hat{v}}_k) \right]$$

where  $k = 1, 2, 3, \dots, R$ , and  $\underline{\hat{v}}_k$  represents the vector of expected values predicted from the "Left Side" for combination  $k$ . In the event that multivariate normality cannot be assumed, Fix and Hodges (1951) utilize a nonparametric procedure which should be examined.

Significance tests can be used to decide whether a significant linear relationship exists between the two sets of variables. It should be noted that, since discriminant analysis may be regarded as a special case of canonical correlation, significance tests for canonical variate pairs will closely resemble those employed for discriminant analysis. The first step in describing the procedure used in overall significance testing is to define Wilk's  $\Lambda$  criterion.

$$\Lambda = \prod_{i=1}^q (1 - \lambda_i^2)$$

where  $q < p$  and where  $\Lambda$  is a statistic which is inversely related to the strength of relationship: the smaller the value of  $\Lambda$ , the greater the relationship strength. After  $\Lambda$  has been computed from the previous equation, an overall significance test may be carried out on the canonical variate pairs by using the chi-square approximation ( $\chi^2$ ). This  $\chi^2$  approximation for the distribution of  $\Lambda$  will provide a test for the null hypothesis that  $p$  variates are not related to the  $q$  variates. The following equality relates  $\chi^2$  to  $\Lambda$ :

$$\chi^2 = -[N - .5(p+q+1)] \ln \Lambda$$

with  $pg$  degrees of freedom and where  $N$  is the total sample size. If we reject the null hypothesis, we can remove the contribution of the first root of  $\Lambda$  and then test the significance of  $q-1$  roots as follows:

$$\Lambda' = \prod_{i=2}^{\min(p,q)} (1 - \lambda_i^2),$$

$$\chi^2 = -[N - .5(p+q+1)] \ln \Lambda' \text{ which has } (p-1)(q-1) \text{ degrees of freedom.}$$

The general equation used for  $r$  roots removed is

$$\Lambda' = \prod_{i=r+1}^{\min(p,q)} (1 - \lambda_i^2)$$

Where  $\chi^2$  is distributed so that there are  $(p-r)(q-r)$  degrees of freedom.

It was originally thought that only the quantity  $\lambda_1^2$  and the corresponding canonical correlation  $R = \lambda_1$  were of any interest. Further examination has shown that, depending upon the research question, roots other than  $\lambda_1^2$  may be relevant. It has been found that one or more subsets of the predictor variables may be related to one or more of the respective subsets of the criterion or predictand variables. The combination of variables in the predictor set  $y$  that are related to a predictand subset in  $x$  can be determined if we inspect the elements of the two vectors  $a_i$  and  $b_i$  which are associated with the quantity  $\lambda_i^2$ . We know that each  $\lambda_i^2$  will be equal to the correlation between the linear functions of the right and left variables, which are formed by using  $b_i$  and  $a_i$  respectively. The chi-square approximation tests that have been defined will show how many of the functions allow statistical interpretation.

As previously discussed, the insurance case study under examination involved a sample size of approximately 20,000 U. S. policies. This particular case study utilized the following two sets of variables:



"Left Side" Variables

- $X_1$  = 1 if policy whole life continuous pay, 0 otherwise.  
 $X_2$  = 1 if policy whole life limited pay, 0 otherwise.  
 $X_3$  = 1 if policy modified life, 0 otherwise.  
 $X_4$  = 1 if policy endowment or retirement, 0 otherwise.  
 $X_5$  = 1 if policy level term, 0 otherwise.  
 $X_6$  = 1 if policy decreasing term, 0 otherwise.  
 $X_7$  = 1 if policy family plan or combination, 0 otherwise.  
 $X_8$  = 1 if policy others with term, 0 otherwise.  
 $X_9$  = 1 if policy > \$50,000., 0 otherwise.  
 $X_{10}$  = 1 if policy \$25,001-\$50,000., 0 otherwise.  
 $X_{11}$  = 1 if policy \$10,001-\$25,000., 0 otherwise.  
 $X_{12}$  = 1 if policy \$10,000., 0 otherwise.  
 $X_{13}$  = 1 if policy less than \$10,000., 0 otherwise.

"Right Side" Variables

- $Y_1$  = 1 if male, 0 otherwise  
 $Y_2$  = 1 if female, 0 otherwise.  
 $Y_3$  = 1 if 15-19 years old, 0 otherwise.  
 $Y_4$  = 1 if 20-24 years old, 0 otherwise.  
 $Y_5$  = 1 if 25-29 years old, 0 otherwise.  
 $Y_6$  = 1 if 30-39 years old, 0 otherwise.  
 $Y_7$  = 1 if single, 0 otherwise.  
 $Y_8$  = 1 if married, 0 otherwise.  
 $Y_9$  = 1 if divorced, widowed, separated, 0 otherwise.  
 $Y_{10}$  = 1 if income < \$3000., 0 otherwise.  
 $Y_{11}$  = 1 if income \$3000-4999., 0 otherwise.  
 $Y_{12}$  = 1 if income \$5000-7499., 0 otherwise.  
 $Y_{13}$  = 1 if income \$7500-9999, 0 otherwise.  
 $Y_{14}$  = 1 if income \$10,000-24,999., 0 otherwise.  
 $Y_{15}$  = 1 if income \$25,000. or over, 0 otherwise.  
 $Y_{16}$  = 1 if not gainfully employed, 0 otherwise.  
 $Y_{17}$  = 1 if occupation professional, 0 otherwise.  
 $Y_{18}$  = 1 if occupation semiprofessional, 0 otherwise.  
 $Y_{19}$  = 1 if student, 0 otherwise.  
 $Y_{20}$  = 1 if housewife, 0 otherwise.  
 $Y_{21}$  = 1 if all others, 0 otherwise.

The sample under examination allowed the following *a priori* probabilities to be estimated:

#### ESTIMATED A PRIORI DISTRIBUTION

	POLICY AMOUNT VS POLICY TYPE							
	WLCP	WLLP	MODL	ENDR	LEVT	DECT	COMB	OTHR
> 50K	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.01
25-50K	0.02	0.00	0.00	0.01	0.01	0.02	0.01	0.01
10-25K	0.05	0.02	0.01	0.02	0.02	0.03	0.04	0.02
10K	0.10	0.04	0.02	0.03	0.01	0.01	0.01	0.03
< 10K	0.20	0.15	0.03	0.04	0.00	0.00	0.00	0.04
TOTAL	0.37	0.21	0.06	0.09	0.04	0.06	0.07	0.10

Three examples obtained by applying the results of the analysis are presented as follows:

Example 1 Buyer - The buyer is a single female, 40 or older, with an income of \$3,000-4999. and she is considered to be a semiprofessional.

#### ESTIMATED PROBABILITIES

	POLICY AMOUNT VS POLICY TYPE							
	WLCP	WLLP	MODL	ENDR	LEVT	DECT	COMB	OTHR
> 50K	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25-50K	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
10-25K	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.00
10K	0.03	0.03	0.01	0.01	0.00	0.00	0.01	0.01
< 10K	0.25	0.44	0.09	0.05	0.00	0.00	0.01	0.04
TOTAL	0.29	0.47	0.10	0.06	0.00	0.01	0.03	0.05

In the above example, the predicted conditional distribution clearly indicates several things. It indicates that there is only a 9% probability of this buyer purchasing a term type policy or any policy combination involving term. This example also indicates that a buyer within this category shows only a 13% probability of purchasing any policy greater than \$10,000. In fact, the modal value of the distribution is 44% that, if a buyer in this category buys, it will be a whole life-limited pay policy under \$10,000.

Example 2 Buyer - This is a married male buyer between twenty and twenty four years of age. This buyer has an annual income of \$25,000. or greater and is not considered a professional or a semi-professional.

#### ESTIMATED PROBABILITIES

	POLICY AMOUNT VS POLICY TYPE							
	WLCP	WLLP	MODL	ENDR	LEVT	DECT	COMB	OTHR
> 50K	0.01	0.00	0.00	0.01	0.00	0.01	0.00	0.01
25-50K	0.05	0.00	0.01	0.03	0.02	0.03	0.02	0.03
10-25K	0.12	0.03	0.01	0.04	0.04	0.07	0.06	0.06
10K	0.13	0.02	0.01	0.03	0.01	0.02	0.01	0.04
< 10K	0.05	0.02	0.00	0.01	0.00	0.00	0.00	0.02
TOTAL	0.35	0.07	0.03	0.11	0.08	0.12	0.09	0.16

The resolution of the combination between size of policy and type of policy is not as distinct as was the case with example 1. The modal value of the distribution is only 13%, that if the potential buyer purchases a policy, it will be a whole life-continuous pay policy equal to \$10,000. Additionally, the predicted distribution shows that there is a 25% probability of a buyer in this

category, if he were to buy, purchasing a whole life-continuous pay type policy in the range of \$10,000. -25,000. . In fact, there is close to a 70% probability that, if a prospect purchases a policy, it will be over \$10,000. .

Example 3 Buyer - This buyer is a male professional who is married and between the ages of thirty and thirty nine and who has an annual income of between \$10,000. and \$24,999. .

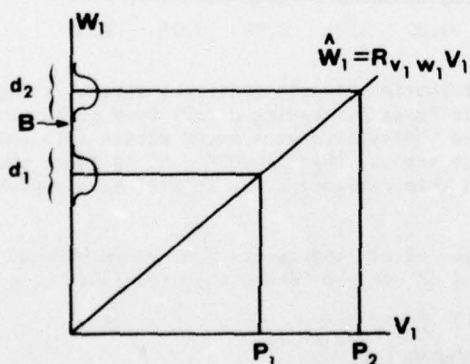
#### ESTIMATED PROBABILITIES

	POLICY AMOUNT VS POLICY TYPE								TOTAL
	WLCP	WLLP	MODL	ENDR	LEVT	DECT	COMB	OTHR	
> 50K	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.03
25-50K	0.04	0.00	0.00	0.02	0.02	0.02	0.02	0.02	0.15
10-25K	0.10	0.03	0.01	0.03	0.04	0.06	0.07	0.04	0.38
10K	0.12	0.03	0.01	0.03	0.01	0.02	0.01	0.03	0.24
< 10K	0.11	0.04	0.01	0.02	0.00	0.00	0.00	0.03	0.21
TOTAL	0.38	0.10	0.03	0.10	0.07	0.10	0.11	0.12	1.00

The conditional probability distribution of Example 3 is similar to that shown by example 2, however, there is a downward shift in policy size. This is no doubt due to the difference in annual income between the Example 2 buyer and the Example 3 buyer. The modal value of Example 3's distribution is 12%, that if the person buys, he will purchase a whole life-continuous pay policy with a value of \$10,000. .

An illustration will now be given of how the above probability tables were obtained. Several dimensions were used above, however, only one will be used in the illustration:

Since we know that  $\lambda_1^2 = R_{V_1 W_1}^2$  gives the maximum correlation between two variates, we can actually draw a regression line whose equation is  $\hat{W}_1 = R_{V_1 W_1} V_1$  where the slope of the line is a function of  $R_{V_1 W_1}$ .



From this line we may obtain the distribution of buyers ( $d_1$ ) for a given product  $P_1$  (40 products may be obtained out of the table) Likewise, we may obtain  $d_2$  for product  $P_2$ . Then we obtain the probabilities of  $P_1$  and  $P_2$  for the buyer B on the  $W_1$  axis, as shown, using Bayes' Theorem.

The above examples indicate the importance of canonical correlation methods as applied to the buyer-product relationship.

Another important application of canonical correlation method deals with the manner in which companies relate to one another based on what they sell and to whom they sell. Given an insurance company's characteristics such as company size, whether it is stock or mutual, ordinary or combination, we can determine the company's peers... insofar as how the company characteristics relate to buyers and products sold.

The same sample was used in this analysis as was used to determine the buyer-product probabilities. Some of the designated variables used in the analysis are as follows:

"Left Side" Variables  
Buyer and Product

$X_1$  = Sex of buyer

"Right Side" Variables  
Company Characteristics

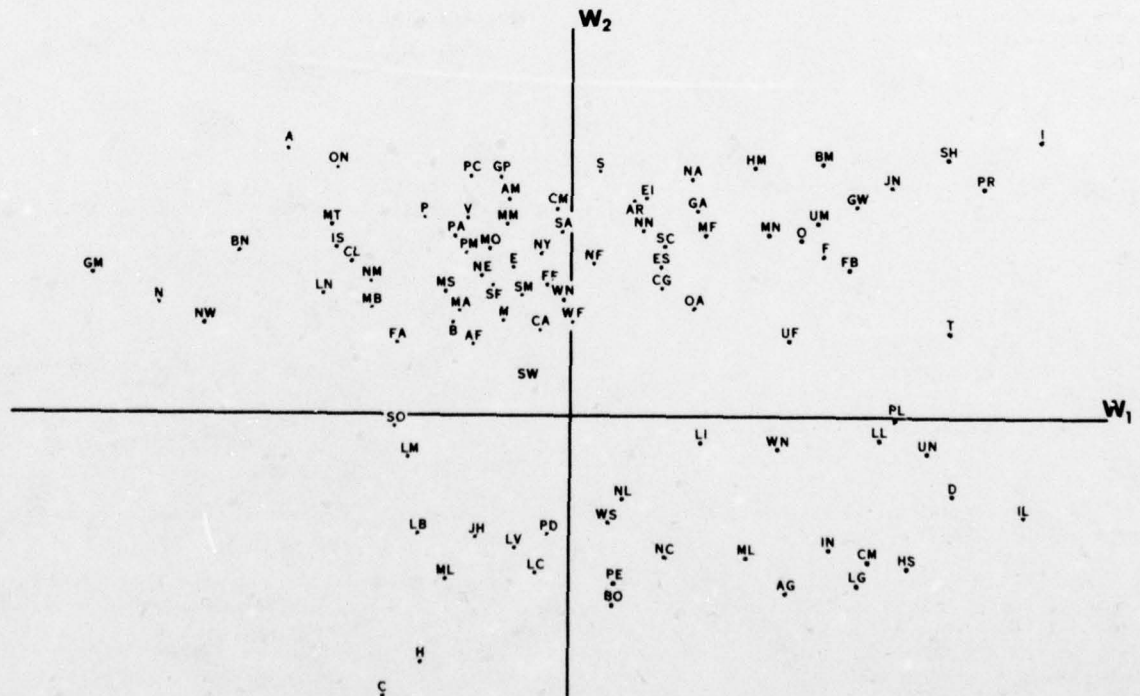
$Y_1$  = Type of company



$X_2$  = Age of buyer  
 $X_3$  = Marital status  
 $X_4$  = Occupation  
 $X_5$  = Income  
 $X_6$  = State of residence  
 $X_7$  = Type of policy  
 $X_8$  = Size of policy  
 $X_9$  = Mode of payment  
 $X_{10}$  = Annualized premium

$Y_2$  = Age of company  
 $Y_3$  = Volume of business  
 $Y_4$  = Area of operation  
 $Y_5$  = Compensation plan  
 $Y_6$  = Commissions paid  
 $Y_7$  = Advertising expenses  
 $Y_8$  = Reserves  
 $Y_9$  = Claims  
 $Y_{10}$  = Participating or nonparticipating  
 $Y_{11}$  = Assets  
 $Y_{12}$  = Licensed in New York  
 $Y_{13}$  = Lapse rate

The following represents a graph of the first two right hand canonical variates  $W_1$  and  $W_2$ . It shows the plotted locations of 94 insurance companies that contributed to the 1970 LIMRA Buyer Study.



A plot, in the  $W_1W_2$  space, of each of the 94 contributors to LIMRA's 1970 Buyer Study. The location of each company point is a function of the company's characteristics. Each characteristic is weighted according to how it relates to who buys company products and what is bought.

The  $W_1$  dimension indicates that the buyers and products sold are related to the size of the company and whether it is a stock or mutual company. In  $W_2$ , which is the second most significant dimension, an ordinary company versus a combination company relationship with buyers and products sold is indicated. From examining the results of this analysis it would seem that more homogeneous comparisons of buyer and product can be made by grouping companies on a size and type basis. To determine the peers for a particular company, it would be appropriate to employ all canonical variates in order to calculate Mahalanobis' distance (the weighted distance between two points).

#### Discussion

Canonical correlation analysis may be regarded as a type of principal components analysis. Thus, the rules by which canonical variates are interpreted are the same as those used to interpret discriminant functions and principal components. The relative magnitudes and signs of several combining weights, which define each of the canonical variates, are closely examined to see if a meaningful interpretation can be given. This type of analysis can have widespread applications. For example, canonical correlation analysis could prove highly useful in studying the relationships between personality attributes and favorable career fields. It would be a means by which to indicate "the right job for the right person." An application such as this could increase job satisfaction among workers and save a company money by decreasing employee turnover due to misplaced individuals.

The above example relates one area in which probability estimates could be formed by using canonical correlation methods. The probabilistic structure of canonical correlation indicates that the use of this type of statistical procedure can and should be further extended into the areas of education, psychology, science and industry. It would seem to have numerous possible applications in meteorology as well.

## Chapter 7

### MARKOV PROCESSES

BY

CAPTAIN ROGER WHITON

#### 1. Introduction.

Many mathematical idealizations or models of nature, as well as a fair number of actual physical processes themselves, have the property that the outcome of any trial or event opportunity depends only on the outcome of the immediately preceding trial. There is no dependency on the history of earlier trials. Processes having this property are referred to as Markov processes or Markov chains, after Andrei Andreevich Markov (1856-1922), whose 1906-1907 studies of the Brownian motion of gas molecules in a closed container laid the groundwork for this subject and later led to the investigation of dependence and stochastic processes. The first correct mathematical construction of a Markov process with continuous trajectories is attributed to Norbert Wiener in 1923. The general theory of Markov processes was developed in the 1930s and 1940s by A.N. Kolmogorov, W. Feller, W. Doeblin, P. Levy, J. L. Doob, and others. Formulation of the Markov chain antedated the principal mathematical development of stochastic processes. In retrospect it can be seen that the Markov chain is actually an extremely simple, nevertheless elegant member of the class of stochastic processes.

It is not difficult to imagine circumstances in which a Markov chain can apply. A classical albeit somewhat artificial example is the random walk problem in which a person or object moves in single steps from one position to another along a line, with one move permitted in each trial. For each current position  $i$ , there exist conditional probabilities for a move from  $i$  to  $i + 1$ , from  $i$  to  $i - 1$ , or from  $i$  to  $i$  (a "stationary move"). The future move is made in accordance with those conditional probabilities, but the determination of which probabilities apply is a function strictly of the earlier trial, which determined the move to  $i$ . Under this conceptualization, the history of moves preceding the one that put the body at  $i$  have exactly no bearing on the move now at hand. Other examples, involving branching problems, multinomial trials, success run chains, certain urn models, and mathematical diffusion models have been developed. With varying degrees of success, the Markov concept has been applied to real phenomena such as paths of free electrons in crystals, queuing problems and even brand selection preferences. Many natural phenomena, however, exhibit complex dependencies and periodicities that tend to violate the simple Markov requirement. Whether the Markov chain is a usefully valid model of these real phenomena depends on the extent to which real world complexities violate the Markov concept and degrade its predictions. In general, the suitability of the Markov chain as a model of real phenomena is determined by empirical test.

The weather has been considered to change from one state or condition to another according to the Markov process, with the outcome of trial (or weather observation)  $n$  depending exclusively on the outcome of the trial or observation  $n - 1$  immediately preceding.

Before undertaking a discussion of Markov chains, let us review some simple concepts from matrix algebra and probability theory.

#### 2. Matrix Multiplication.

If  $\vec{A}$  is an  $m \times p$  matrix ( $m$  = number of rows  $i$  and  $p$  = number of columns  $j$ ), and if  $\vec{B}$  is a  $p \times n$  matrix, the product  $\vec{C} = \vec{A} \vec{B}$  of the two matrices exists. Such a product must be a  $m \times n$  matrix. The rule for formation of the matrix product is

$$c_{ij} = \sum_{k=1}^p a_{ik} b_{kj}$$

If  $\vec{A}$  is an  $n$ -square matrix, then we can form all the powers of  $\vec{A}$ , namely:

$$\vec{A}^2 = \vec{A} \vec{A}$$

$$\vec{A}^3 = \vec{A} \vec{A}^2$$

...

Routines for matrix multiplication are normally resident in the mathematics libraries of computer installations and are useful in the case of matrices larger than  $3 \times 3$  or for chains of matrix multiplications. Smaller problems are easily handled by manual methods.

#### 3. Matrix Inversion.

If  $\vec{A}$  is a non-singular,  $m \times m$ , square matrix, it has an inverse  $\vec{A}^{-1}$  such that



$$\vec{A} \vec{A}^{-1} = \vec{I}$$

where  $\vec{I}$  is the identity matrix (consisting of all ones along the main diagonal).

The process of inverting a matrix is inherently more complex than that of matrix multiplication. Computer library matrix inversion routines are correspondingly more valuable and more often used than routines for matrix multiplication. While a variety of matrix inversion methods is available, a simple technique usable for hand inversion of smaller matrices (up to 3 x 3 conveniently) is illustrated below. This is not usually the method selected for computer implementation.

In the inversion of an  $m \times m$  matrix  $\vec{A}$  by the method of cofactors, one first forms the cofactor matrix  $\vec{A}^c$  of  $\vec{A}$ . Each element  $a_{ij}^c$  of the cofactor matrix is the determinant of the minor formed by deleting from the original matrix the row  $i$  and column  $j$ . For example,

$$\begin{aligned} a_{11}^c &= a_{22}a_{33} - a_{23}a_{32} \\ a_{12}^c &= a_{21}a_{33} - a_{23}a_{31} \\ &\dots \quad \dots \quad \dots \end{aligned}$$

With the cofactor matrix  $\vec{A}^c$  available, the next step is to form the so called adjoint matrix  $\vec{A}^a$ , which is simply the transpose of the cofactor matrix  $\vec{A}^c$ :

$$\vec{A}^a = (\vec{A}^c)'$$

Transposing a matrix is simple. One writes as the columns of the transpose those elements comprising the rows of the original. Thus,

$$\vec{A}^a = (\vec{A}^c)' = \begin{bmatrix} a_{11}^c & a_{21}^c & a_{31}^c \\ a_{12}^c & a_{22}^c & a_{32}^c \\ a_{13}^c & a_{23}^c & a_{33}^c \end{bmatrix}$$

With the adjoint matrix available, one forms the inverse of  $\vec{A}$  simply by dividing the adjoint matrix  $\vec{A}^a$  by the determinant  $|\vec{A}|$  of the original matrix, i.e.,

$$\vec{A}^{-1} = \vec{A}^a / |\vec{A}|$$

Since  $|\vec{A}|$  is a scalar, one simply divides each element  $a_{ij}^a$  of the adjoint by that number to obtain the inverse  $\vec{A}^{-1}$ . This manual technique has been used to invert the matrices in this chapter because they are all 3 x 3 or smaller. Technically, the method is referred to as Cramer's rule.

#### 4. Fixed Vector or Fixed Point.

Consider the vector (row matrix)  $\vec{q}$  having  $n$  components, where

$$\vec{q} = [q_1 \quad q_2 \quad q_3 \quad \dots \quad q_j \quad \dots \quad q_{n-1} \quad q_n]$$

Such a vector is  $(1 \times n)$ -dimensional. Provided the matrix  $\vec{A}$  is square and  $(n \times n)$ , we can define a  $(1 \times n)$  matrix product  $\vec{q} \vec{A}$ . If, furthermore,

$$\vec{q} \vec{A} = \vec{q}$$

then we say that  $\vec{q}$  is "left fixed" (not changed) by its multiplication with  $\vec{A}$ . Any vector  $\vec{q} \neq 0$  is a fixed vector (or so called fixed point) of  $\vec{A}$  if it is left fixed when multiplied by  $\vec{A}$ . To illustrate, let us choose

$$\vec{q} = [q_1 \quad q_2] \quad (n = 2)$$

$$\vec{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad (m = n = 2)$$

$$\vec{q} \vec{A} = [(q_1 a_{11} + q_2 a_{21}) \quad (q_1 a_{12} + q_2 a_{22})]$$

If  $\vec{q} \vec{A} = \vec{q}$ , then

$$q_1 a_{11} + q_2 a_{21} = q_1$$

$$q_1 a_{12} + q_2 a_{22} = q_2$$

$$q_2 = q_1 \frac{a_{12}a_{11}}{1 - a_{22} - a_{12}a_{21}}$$

If we are given, for example, the matrix,

$$\vec{A} = \begin{bmatrix} 2 & 1 \\ 2 & 3 \end{bmatrix}$$

then we can develop any number of fixed vectors  $\vec{q}$  of the matrix  $\vec{A}$  simply by choosing arbitrary  $q_1$ . For example, if we select  $q_1 = -1$  and  $q_2 = 1/2$ , giving

$$\vec{q} = \begin{bmatrix} -1 & 1/2 \end{bmatrix}$$

If, on the other hand, we select  $q_1 = 2$ , then

$$\vec{q} = \begin{bmatrix} 2 & -1 \end{bmatrix}$$

These are two of infinitely many fixed vectors of  $\vec{A}$

##### 5. Probability Vectors and Stochastic Matrices.

The probabilities associated with various states of a system may be expressed in terms of a probability vector  $\vec{p}$  having one element for each such state. For example, if the weather is characterized in terms of three mutually exclusive and exhaustive states, "stormy," "unsettled" and "fair" (states 1, 2 and 3, respectively), then the likelihood of occurrence of each state is given by a probability vector  $\vec{p}$ :

$$\vec{p} = \begin{bmatrix} p_1 & p_2 & p_3 \end{bmatrix}$$

We say that a vector,

$$\vec{p} = \begin{bmatrix} p_1 & p_2 & p_3 & \dots & p_j & \dots & p_{n-1} & p_n \end{bmatrix}$$

is a probability vector if its components are non-negative and their sum is unity.

Extending this idea of probability as a vector, where the vector is a one-dimensional matrix, we can express the likelihood of transition from one state to another as a stochastic matrix. For example, in the three-state problem given above, if the present weather is "unsettled" (state 2), then it can either change to "fair" (a 2,3 transition), change to "stormy" (a 2, 1 transition), or remain the same (a 2,2 "transition"). Thus, associated with present state 2 there are three transition probabilities:  $p_{21}$ ,  $p_{22}$  and  $p_{23}$ . Likewise, there are three other transition probabilities associated with present state 1 and three more with present state 3. Plainly, the likelihood of "change" in the weather can in our three-state problem be characterized in terms of a  $3 \times 3 = 9$ -component matrix:

$$\vec{P} = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix}$$

As it turns out, such a matrix is a stochastic matrix. In fact, it is a particular form of stochastic matrix called a "transition matrix" which will be described in a later section. The probabilities  $p_{ij}$  are conditional probabilities that state  $j$  will occur given the system is in state  $i$ . In equivalent notation,

$$p_{ij} = P \{ a_j \mid a_i \}$$

A square matrix  $\vec{P} = [p_{ij}]$  is called a stochastic matrix if each of its rows is a probability vector. If two matrices  $\vec{P}_1$  and  $\vec{P}_2$  are stochastic, their product  $\vec{P}_1 \vec{P}_2$  and all the powers  $\vec{P}^n$  and  $\vec{P}_2^n$  are also stochastic matrices. A stochastic matrix  $\vec{P}$  is said to be regular if all the elements of any of its powers  $\vec{P}^n$  are positive. Zeroes or negative numbers are disqualifying.

Regular stochastic matrices have mathematically attractive properties. If  $\vec{P}$  is a regular stochastic matrix, then it follows that:

- Associated with  $\vec{P}$  is a unique fixed probability vector  $\vec{t}$  each of whose components is positive and for which, by definition,

$$\vec{t} \vec{P} = \vec{t}$$

\*We are using "change" and "transition" in their extended sense, which includes the act of remaining the same, or "persisting."

- The sequence of powers of  $\vec{P}$ , namely,  $\vec{P}, \vec{P}^2, \vec{P}^3, \dots$ , approaches a matrix  $\vec{T}$  each of whose rows is simply the fixed probability vector  $\vec{t}$ .

- If  $\vec{a}$  is any probability vector, then the sequence of vectors  $\vec{a}\vec{P}, \vec{a}\vec{P}^2, \vec{a}\vec{P}^3, \dots$  approaches the unique fixed probability vector  $\vec{t}$ .

Notice we said the fixed vector  $\vec{t}$  is "unique," whereas fixed vectors in general are far from unique. What makes  $\vec{t}$  unique is our insistence that it be more than an ordinary fixed vector. We required that it also be a probability vector, the sum of whose elements is unity. This additional constraint is sufficient to reduce the infinity of candidate fixed vectors of  $\vec{P}$  to a single fixed probability vector  $\vec{t}$ , in other words, a unique fixed vector.

To see how this works, let us consider the transition matrix  $\vec{P}$  characterizing the three-state weather problem discussed earlier:

$$\vec{P} = \begin{bmatrix} 0.4 & 0.5 & 0.1 \\ 0.1 & 0.5 & 0.4 \\ 0.01 & 0.09 & 0.9 \end{bmatrix}$$

We seek a fixed probability vector,

$$\vec{t} = [t_1 \quad t_2 \quad (1-t_1-t_2)]$$

such that

$$\vec{t}\vec{P} = \vec{t}$$

Performing the indicated matrix multiplication  $\vec{t}\vec{P}$  and setting the product equal to the vector  $\vec{t}$  yields the system,

$$p_{31} + (p_{11}-p_{31}-1)t_1 + (p_{21}-p_{31})t_2 = 0$$

$$p_{32} + (p_{12}-p_{32})t_1 + (p_{22}-p_{32}-1)t_2 = 0$$

$$p_{33} - 1 + (p_{13}-p_{33}+1)t_1 + (p_{23}-p_{33}+1)t_2 = 0$$

We can simplify the notation by using

$$\begin{array}{lll} a = p_{31} & d = p_{32} & g = p_{33} - 1 \\ b = p_{11}-p_{31}-1 & e = p_{12}-p_{32} & h = p_{13}-p_{33}+1 \\ c = p_{21}-p_{31} & f = p_{22}-p_{32}-1 & i = p_{23}-p_{33}+1 \end{array}$$

which yields

$$a + bt_1 + ct_2 = 0$$

$$d + et_1 + ft_2 = 0$$

$$g + ht_1 + it_2 = 0$$

where obviously the final equation is superfluous. The problem reduces to one of solving two simultaneous equations. From the first,

$$t_2 = -\frac{a + bt_1}{c}$$

Using the second,

$$t_1 = \frac{fa - cd}{ce - fb}$$

Substituting the numerical values provided in the transition matrix, we obtain

$$\begin{array}{lll} a = 0.01 & d = 0.09 & g = -0.10 \\ b = -0.61 & e = 0.41 & h = 0.20 \\ c = 0.09 & f = -0.59 & i = 0.50 \end{array}$$



from these values, we obtain

$$t_1 = 0.0433$$

and

$$t_2 = 0.1827$$

Therefore,

$$t_3 = 1 - t_1 - t_2 = 0.7740$$

Hence,

$$\vec{t} = [0.0433 \quad 0.1827 \quad 0.7740]$$

That  $\vec{t}$  is indeed a fixed vector of  $\vec{P}$  can be verified by performing the matrix multiplication  $\vec{t}\vec{P}$  and noting that the product is equal to  $\vec{t}$  within the limits of truncation error.

When the dimensionality, i.e., the number of Markov states, of the problem is large, the straightforward algebraic method shown above for obtaining the fixed probability vector  $t$  becomes unwieldy. Under these circumstances, matrix solutions may prove advantageous, since computational routines for them are available in many computer libraries. Let us reconsider our set of simultaneous equations:

$$bt_1 + ct_2 = -a$$

$$et_1 + ft_2 = -d$$

or

$$\begin{bmatrix} b & c \\ e & f \end{bmatrix} \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} = \begin{bmatrix} -a \\ -d \end{bmatrix}$$

Hence,

$$\begin{bmatrix} t_1 \\ t_2 \end{bmatrix} = \begin{bmatrix} b & c \\ e & f \end{bmatrix}^{-1} \begin{bmatrix} -a \\ -d \end{bmatrix}$$

Taking the inverse,

$$\begin{aligned} \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} &= \begin{bmatrix} f/(bf-ce) & -c/(bf-ce) \\ -e/(bf-ce) & b/(bf-ce) \end{bmatrix} \begin{bmatrix} -a \\ -d \end{bmatrix} \\ &= \begin{bmatrix} (cd-fa)/(bf-ce) \\ (ae-bd)/(bf-ce) \end{bmatrix} \end{aligned}$$

and

$$\vec{t} = [0.0433 \quad 0.1827 \quad 0.7740]$$

which was the result obtained previously by non-matrix methods.

One of the properties of the regular stochastic matrix  $\vec{P}$  is that the sequence of its powers approaches the matrix  $\vec{T}$  each of whose rows is simply the fixed probability vector  $\vec{t}$  associated with  $\vec{P}$ . In the present case,

$$\vec{T} = \begin{bmatrix} 0.0433 & 0.1827 & 0.7740 \\ 0.0433 & 0.1827 & 0.7740 \\ 0.0433 & 0.1827 & 0.7740 \end{bmatrix}$$

We can illustrate this by forming some of the powers of the matrix  $\vec{P}$ :

$$\vec{P}^2 = \vec{P}\vec{P} = \begin{bmatrix} 0.2110 & 0.4590 & 0.3300 \\ 0.0940 & 0.3360 & 0.5700 \\ 0.0220 & 0.1310 & 0.8470 \end{bmatrix}$$

$$\vec{p}^3 = \vec{p}^2 \vec{P} = \begin{bmatrix} 0.1336 & 0.3647 & 0.5017 \\ 0.0769 & 0.2663 & 0.6568 \\ 0.0304 & 0.1527 & 0.8169 \end{bmatrix}$$

$$\vec{p}^6 = \vec{p}^5 \vec{P} = \begin{bmatrix} 0.0611 & 0.2225 & 0.7164 \\ 0.0507 & 0.1993 & 0.7500 \\ 0.0406 & 0.1765 & 0.7829 \end{bmatrix}$$

$$\vec{p}^{12} = \vec{p}^{11} \vec{P} = \begin{bmatrix} 0.0441 & 0.1844 & 0.7715 \\ 0.0437 & 0.1834 & 0.7730 \\ 0.0432 & 0.1824 & 0.7744 \end{bmatrix}$$

Eventually,  $\vec{p}^{22}$  agrees with  $\vec{T}$  to the fourth decimal place.

The fact that the transition matrix  $\vec{P}$  converges to  $\vec{T}$  in sufficiently great powers gives the fixed probability vector  $\vec{T}$  a special meaning in the Markov process. As shown below,  $\vec{T}$  represents the long term likelihood of occurrence of each of the Markov states over many Markov "trials." Meteorologically, the  $\vec{T}$  vector embodies the unconditional probabilities  $p_j = p\{a_j\}$  of each of the Markov states, i.e., the climatological relative frequency of the states.

#### 6. The Markov Process or Markov Chain.

The Markov chain is a mathematical model of the behavior of a system. When used to represent actual systems, real processes or the behavior of nature, the Markov chain represents a simplified generalization of complex, varied reality. Like any model, the Markov chain succeeds to a greater or a lesser extent in portraying the actual behavior of real systems, depending on the extent to which those systems correspond to the requirements of the model.

The Markov model of the behavior of a system has two such requirements or defining properties:

- The system under consideration may be categorized as being in one of a finite number of states  $a_i$ , the complete set of which, namely,

$$\{a_1 \quad a_2 \quad a_3 \quad \dots \quad a_i \quad \dots \quad a_{m-1} \quad a_m\}$$

constitutes the state space of the system.

- At each trial, the system has the opportunity either to change its state or to remain in the same state. The outcome of any trial in terms of the state of the system depends at most upon the outcome of the immediately preceding trial, and not upon any previous outcome.

The Markov process or so called finite Markov chain is a stochastic process embodying these two model conditions. In the Markov process, we envision the state of the system changing from  $a_i$  to  $a_j$ , where either  $i \neq j$  or  $i = j$ . This is termed an  $(a_i, a_j)$ -transition, or in short an  $(i, j)$ -transition meaning the state  $a_j$  occurs immediately after  $a_i$  occurs. In other words, the outcome of trial  $n$  is state  $a_i$  and the outcome of trial  $n+1$  is state  $a_j$ . The probability that a system in state  $a_i$  will undergo a transition to state  $a_j$  is  $p_{ij}$ , called a transition probability, a conditional probability. As we discussed before, the transition probabilities  $p_{ij}$  form a transition matrix  $\vec{P}$ , an  $m \times m$  stochastic matrix where  $m$  is the number of permitted states  $a_i$ .

$$\vec{P} = \begin{bmatrix} p_{11} & p_{12} & \dots & p_{1m} \\ p_{21} & p_{22} & \dots & p_{2m} \\ \dots & \dots & p_{ij} & \dots \\ p_{m1} & p_{m2} & \dots & p_{mm} \end{bmatrix} \quad \text{where} \quad \begin{array}{l} i = \text{Current state} \\ j = \text{Future state} \end{array}$$

For each current state  $a_i$ , the  $i$ th row of the transition matrix is the conditional probability vector of all possible state outcomes in the next trial. The fact that  $\vec{P}$  is a regular stochastic matrix guarantees each of its rows will be a probability vector.

Markov models have been used to attempt to represent changes in the weather as transitions from one Markov state to another. These meteorological applications of the Markov model consider nature has a new "trial" at each of  $N$  regularly spaced observation intervals  $\Delta t$ , where  $\Delta t$  might logically be the weather observation interval of 1 hr. For each such trial, there must be an outcome in the form of a Markov state  $a_j$ .

We can illustrate by categorizing the behavior of the weather in terms of the same three Markov states we considered before:

$a_1$  = Stormy

$a_2$  = Unsettled

$a_3$  = Fair

This three-state Markov problem is an oversimplification used here for illustrative purposes. In actual problems, in order to obtain needed resolution in the forecast scheme, there can be hundreds of Markov states corresponding to realistically resolved discretizations of such observed variables as ceiling, visibility, sea level pressure, temperature, dew point temperature, wind, and others. Methods for handling  $n$ -state Markov chains will be discussed in a later section. Meanwhile, the three-state problem will suffice to exemplify basic Markov concepts.

The foundation upon which any Markov analysis stands is historical data on the performance of the system over many trials. In the meteorological sense, this historical data represents climatology and might be a list of weather observations such as that shown below:

Trial = Time*	Outcome = State of Nature
1	$a_2$
2	$a_1$
3	$a_3$
4	$a_3$
5	$a_2$
...	...

If sufficient data are available, the sequence of trials and outcomes called climatology can be converted into a Markov transition matrix  $\vec{P}$ . For convenience, let us suppose our data have yielded the transition matrix discussed in section 5 above. The interpretation of the  $\vec{P}$  matrix is clear. Let us imagine that the current state is "unsettled" ( $a_2 = 1, i = 2$ ). Under these circumstances, the applicable vector of conditional probabilities is the second row of the  $\vec{P}$  matrix, i.e.,

$$\vec{p}_2 = [0.1 \quad 0.5 \quad 0.4]$$

This probability vector indicates, given that the weather is now "unsettled," there is a 10 percent likelihood conditions will change to "stormy," a 40 percent chance they will improve to "fair," and a 50 percent chance the weather will remain the same. These probabilities can be taken as a probabilistic forecast of the state of the system for one time step in the future, i.e., at time  $t_0 + \Delta t$ , where  $t_0$  is the initial time. For the case where the transition matrix is based on one-hour weather changes, the probability vector provides a one-hour probabilistic weather forecast.

But can we apply the Markov model to weather changes? If we are to do so, the weather must meet the condition that the outcome of any trial or weather observation depends only on the outcome of the preceding trial or observation. It is by no means certain that this is the case. Indeed the notorious dependency and periodicity of weather data suggest such a premise is not true in general. Unlike the classical random walk and coin toss problems, in which the Markov

---

\*Units are arbitrary and problem-dependent. In problems where observations at 1 hr intervals are used to construct the transition matrix,  $\Delta t = 1$  hr, and the unit of time is hours.



requirement is met, or the urn problem without replacement, in which it is not, there are no a priori means of ascertaining whether the weather meets the requirements of a Markov process. On the other hand, the Markov chain may be sufficiently effective as a model of atmospheric behavior to have skill in weather prediction. The best test is to state the process of weather change in Markov terms and then to consider empirically the performance of the Markov model in simulating weather events. If, for example, a long series of Markov predictions neither violates climatology nor lacks skill in prognosis when applied to independent data, then the Markov model can be considered applicable to weather events.

If our discussion were to end here, we might be justified in concluding the Markov concept has little practical value, being limited merely to forecasts for the next trial, i.e., the next observation time. In practice, the meteorologist prepares forecasts for several hours in advance, not just one. Fortunately, the Markov process readily extends to higher order transition probabilities by means of an n-step Markov chain. This concept is discussed in the following section.

#### 7. The n-Step Markov Chain and Higher Order Transition Probabilities.

Assuming the weather changes according to a Markov process in which "trials" occur at intervals of the normal hourly observation time, let us consider the problem of making a three-hour forecast of "stormy," "unsettled" or "fair" in our three-state system, based on an initial state of "unsettled" ( $i = 2$ ).

If our problem had been to make a one-hour forecast, the problem would have been simple. The forecast result is simply the probability vector  $\vec{p}_2$  discussed above. This is the trivial one-step Markov process. Instead of this, our actual problem is the n-step Markov process, in which  $n$  is three in this case.

Specifically, we desire the probability  $p_{ij}^{(n)}$  that the system changes from state  $a_i$  to  $a_j$  in exactly  $n$  steps. In general, the  $n$  states of nature can be given as

$$a_i^{(0)} \rightarrow a_k^{(1)} \rightarrow a_k^{(2)} \rightarrow \dots \rightarrow a_k^{(n-1)} \rightarrow a_j^{(n)}$$

or in our particular three-step case,

$$a_{i=2}^{(2)} \rightarrow a_k^{(1)} \rightarrow a_k^{(2)} \rightarrow a_j^{(3)}$$

The probabilities  $p_{ij}^{(n)}$  of the n-step transition from  $a_i$  to  $a_j$  form what is known as an n-step transition matrix  $\vec{P}^{(n)}$ . The n-step transition matrix turns out to equal the nth power of the original stochastic transition matrix  $\vec{P}$ . Thus,

$$\vec{P}^{(n)} = \vec{P}^n$$

In our three-state problem, we earlier calculated the third power of  $\vec{P}$  as

$$\vec{P}^{(3)} = \begin{bmatrix} 0.1336 & 0.3647 & 0.5017 \\ 0.0769 & 0.2663 & 0.6568 \\ 0.0304 & 0.1527 & 0.8169 \end{bmatrix}$$

For the case of  $i = 2$  initially, the conditional probabilities governing the three-hour forecast would be

$$\vec{p}_2^{(3)} = [0.0769 \quad 0.2663 \quad 0.6568]$$

where the most likely event (with a 66 percent chance) is improvement in the weather. Note how this contrasts with the one-hour forecast, where persistence of "unsettled" weather was most likely:

$$\vec{p}_2 = [0.1 \quad 0.5 \quad 0.4]$$

Since  $\vec{P}$  is a regular stochastic matrix, its powers approach the matrix  $\vec{T}$  each of whose rows is simply the fixed probability vector  $\vec{t}$ . We illustrated this before. For example, the forecast for six hours ( $n = 6$ ) is

$$\vec{p}_2^{(6)} = [0.0507 \quad 0.1993 \quad 0.7500]$$

and for 12 hours ( $n = 12$ ) is

$$\vec{p}_2^{(12)} = [0.0437 \quad 0.1834 \quad 0.7730]$$

At 24 hours we have

$$\vec{p}_2^{(24)} = [0.0433 \quad 0.1827 \quad 0.7740] = \vec{t}$$

This convergence of  $\vec{p}^{(n)}$  to  $\vec{t}$  leads to the subject of the stationary distribution of regular Markov chains.

#### 8. Stationary Distribution of the Markov Chain.

If a Markov problem is characterized by a transition matrix  $\vec{P}$  that is regular, then the sequence of n-step transition matrices  $\vec{P}^n$  approaches the matrix  $\vec{T}$  each of whose rows is the fixed probability vector  $\vec{t}$ . Hence, for sufficiently many Markov steps (i.e., for long range forecasts), the conditional probability  $p_{ij}^{(n)}$  that state  $a_j$  occurs n steps after  $a_i$  becomes independent of the original state  $a_i$  and approaches the component  $t_j$  of  $\vec{t}$ , the fixed probability vector of  $\vec{P}$ .

This leads to the concept that  $\vec{t}$  represents the stationary distribution of the Markov chain, i.e., that distribution of transition probabilities that obtains after a large number of steps of the Markov process. The fixed probability vector  $\vec{t}$ , then, represents the long term likelihood that state space  $a$  will obtain. The likelihood of state  $a_i$  is given by the component  $t_i$  in the long run. In the meteorological sense, the fixed probability vector  $\vec{t}$  represents climatology, the expected long term relative frequency of occurrence of states of the weather.

The fact that the transition matrix of forecast conditional probabilities  $\vec{P}^n$  approaches climatology  $\vec{T}$  as the forecast period  $n \Delta t$  lengthens is highly attractive from the point of view of a practical prediction scheme. Almost any experienced forecaster, when asked for a forecast beyond the period for which conventional prognostic techniques show skill, will turn to the climatology to construct his "prog."

The convergence of  $\vec{P}^n$  to the unchanging matrix of climatology  $\vec{T}$  provides us with a useful intuitive perception of how the Markov process works. We see that, as time goes on, the effect of the initial state on the likelihood of occurrence of future states appears to "wear off." This is a physically reasonable result. We would hardly expect this hour's observation of the visibility at Spokane to have much bearing on the probability distribution of visibility states there a year from now. On the other hand, the likelihood of visibility less than 1 mile there an hour from now would be enhanced greatly if the current observation were to show visibility restrictions.

Thus we can think of the Markov chain as modeling the decorrelation of weather events over time.

#### 9. State Probability Distributions vs. Transition Probabilities.

So far we have considered only the conditional probabilities expressed in the Markov transition matrix. Such a probability might, for example, be  $p_{12}$ , the conditional probability, given the system is in state 1, that it will change to state 2. For forecast purposes, however, it is often more useful to have the forecast probabilities of occurrence of each of the Markov states  $a_j$  for the forecast time  $t = t_0 + n \Delta t$ . We consider in this section how these state probabilities can be obtained.

If  $p_{12}$  is the conditional probability of state 2 at time  $t$  given state 1 at time  $t_0$ , and if  $q_1$  is the probability of state 1 at time  $t_0$ , then the probability of state 2 at  $t$  is simply the product of  $q_1 p_{12}$ . The probability  $q_1$  is called an initial probability. There is actually a distribution of initial probabilities, one for each Markov state, forming an initial probability vector:

$$\vec{q} = [q_1 \quad q_2 \quad q_3]$$

According to the reasoning we used above, the state probability distribution of the system at time  $t$  is simply the product,

$$\vec{q} \vec{P}^{(n)} = \begin{bmatrix} q_1 p_{11}^{(n)} + q_2 p_{21}^{(n)} + q_3 p_{31}^{(n)} \\ q_1 p_{12}^{(n)} + q_2 p_{22}^{(n)} + q_3 p_{32}^{(n)} \\ q_1 p_{13}^{(n)} + q_2 p_{23}^{(n)} + q_3 p_{33}^{(n)} \end{bmatrix}$$

In some applications, particularly climatological ones, the initial probability distribution is fractional and expresses the likelihood  $q_i$  that the system will start in Markov state  $a_i$ . In weather forecasting applications of the Markov model, however, there is usually no uncertainty about

the initial state, since that state is almost always the observation (at time  $t_0$ ) upon which the forecast for time  $t$  is based. Accordingly, the initial probability distribution is generally a vector consisting of all zeroes and a single one. If, for example, the initial state is 2 ("unsettled"), then

$$\vec{q} = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$$

and the forecast state probabilities for time  $t$  reduce to

$$\vec{q} \vec{P}^{(n)} = \begin{bmatrix} p_{21}^{(n)} & p_{22}^{(n)} & p_{23}^{(n)} \end{bmatrix}$$

This result can be verified by performing the matrix multiplication. Note that in obtaining  $p_{ij}^{(n)}$ , the entire matrix  $\vec{P}$  must be raised to the power  $n$ , not the elements separately.

It is apparent that the Markov model can generate a probabilistically expressed forecast of the state of the weather at a future time  $t$  based on an observation of the state at time  $t_0$ . In section 11, we will address the complexity of  $n$ -state Markov models needed to express the state of the weather in all its variety.

#### 10. Eigenvalue Methods for Powers of Matrices.

In both the classical Markov analysis and the equivalent Markov approach to weather forecasting (see section 11 below), obtaining the forecast for time  $t = t_0 + n \Delta t$  involves taking the  $n^{\text{th}}$  power of either the Markov transition matrix  $\vec{P}$  or the equivalent Markov matrix of coefficients  $\vec{B}$  (or  $\vec{A}$ ). For matrices whose dimensionality exceeds three, or for large  $n$ , taking powers of the matrix explicitly is inconvenient. Even some small computers may be insufficiently equipped to do this job efficiently. And in the absence of a computer or tables of powers of the matrix  $\vec{P}$  or  $\vec{B}$ , it becomes difficult to apply Markov techniques to real weather forecasting problems.

Fortunately a speedy, computationally simple means exists for taking powers of a matrix such as  $\vec{P}$  without performing the matrix multiplication explicitly. This method is equally suited to human or computer use and should be resorted to whenever powers of a matrix are desired.

To obtain the  $n^{\text{th}}$  power of any square matrix  $\vec{P}$ , one applies the rule given by Feller (1968, pp 428-432):

$$\vec{P}^n = \lambda_1 \frac{n \vec{x}_1 \vec{y}_1}{\vec{y}_1 \vec{x}_1} + \lambda_2 \frac{n \vec{x}_2 \vec{y}_2}{\vec{y}_2 \vec{x}_2} + \lambda_3 \frac{n \vec{x}_3 \vec{y}_3}{\vec{y}_3 \vec{x}_3} + \dots + \lambda_m \frac{n \vec{x}_m \vec{y}_m}{\vec{y}_m \vec{x}_m}$$

Where  $m$  is the dimensionality of the  $m \times m$  matrix  $\vec{P}$ ,  $\lambda_i$  is the  $i^{\text{th}}$  eigenvalue of that matrix,  $\vec{x}_i$  is the "left" eigenvector associated with  $\lambda_i$ , and  $\vec{y}_i$  is the "right" eigenvector associated with  $\lambda_i$ . By "left" eigenvectors, we mean eigenvectors of the original matrix  $\vec{P}$ . "Right" eigenvectors are eigenvectors of its transpose  $\vec{P}'$ . Since all the left eigenvectors  $\vec{x}$  are column vectors ( $m \times 1$ ), and all the right eigenvectors  $\vec{y}$  are row vectors ( $1 \times m$ ), the products  $\vec{x} \vec{y}$  are  $m \times m$  square matrices, the same size as the original  $\vec{P}$  matrix. On the other hand, the products  $\vec{y} \vec{x}$  are ( $1 \times 1$ ), i.e., scalars. Since the eigenvalues  $\lambda$  and their powers  $\lambda^n$  are also scalars, it is apparent that the problem of taking the  $n^{\text{th}}$  power of  $\vec{P}$  reduces to that of taking the sum of  $m$  different  $m \times m$  matrices, each weighted by  $\lambda_i^n / \vec{y}_i \vec{x}_i$ . Once the eigenvalues  $\lambda_i$  and associated left and right eigenvectors  $\vec{x}_i$  and  $\vec{y}_i$  are found, the process of obtaining  $\vec{P}^n$  by this rule becomes trivial.

Normally, finding eigenvalues and eigenvectors is done on a computer, using library subroutines. To illustrate what the eigenvalues and eigenvectors are, however, we will solve one such problem by hand, finding the eigenvalues and eigenvectors of the Markov transition matrix  $\vec{P}$  presented earlier.

The eigenvalues  $\lambda_i$  (also called characteristic values, characteristic roots, latent roots, proper values or proper numbers) of a matrix  $\vec{P}$  are the roots of the characteristic equation of the matrix. The eigenvalues  $\lambda_i$  are defined such that corresponding to each of them is a non-zero vector  $\vec{x}_i$  called an eigenvector for which

$$\vec{P} \vec{x}_i = \lambda_i \vec{x}_i$$

For  $\vec{P}$  with dimensions  $m \times m$ , there exist  $m$  eigenvalues  $\lambda_i$ . The eigenvalues are scalars. To each eigenvalue there corresponds an eigenvector  $\vec{x}_i$  consisting of a column of  $m$  elements (right eigenvectors, however, are considered row vectors).

From the equation above,

$$\vec{P} \vec{x}_i - \lambda_i \vec{x}_i = 0$$

which is equivalent to



$$(\vec{P} - \lambda_i \vec{I}) \vec{x}_i = 0$$

where  $\vec{I}$  is the identity matrix. This is analogous to the system of simultaneous, homogeneous equations,

$$\vec{A} \vec{x} = 0$$

for whose non-trivial solution we require that the "denominator determinant" be zero, i.e., that the matrix of coefficients  $\vec{A}$  be singular (not have an inverse). Mathematically,

$$\det \vec{A} = |\vec{A}| = 0$$

leads to the solution vector  $\vec{x}$ . In the case of our eigenvalue problem, we require

$$\det (\vec{P} - \lambda_i \vec{I}) = |\vec{P} - \lambda_i \vec{I}| = 0$$

This is called the characteristic equation of the matrix  $\vec{P}$ :

$$|\vec{P} - \lambda_i \vec{I}| = 0$$

The  $m$  eigenvalues  $\lambda_i$  are the roots of the characteristic equation.

Let us expand the determinant and evaluate it:

$$|\vec{P} - \lambda_i \vec{I}| = \begin{vmatrix} 0.4 - \lambda_i & 0.5 & 0.1 \\ 0.1 & 0.5 - \lambda_i & 0.4 \\ 0.01 & 0.09 & 0.9 - \lambda_i \end{vmatrix} = 0$$

which yields the cubic equation,

$$\lambda_i^3 - 1.8 \lambda_i^2 + 0.923 \lambda_i - 0.123 = 0$$

Standard algebraic methods produce from this cubic the three roots,

$$\begin{aligned} \lambda_1 &= 1.0000 \\ \lambda_2 &= 0.5924 \\ \lambda_3 &= 0.2076 \end{aligned}$$

which are the eigenvalues of  $\vec{P}$ . Associated with each eigenvalue  $\lambda_i$  is an eigenvector  $\vec{x}_i$  of  $m$  elements, obtainable from

$$(\vec{P} - \lambda_i \vec{I}) \vec{x}_i = 0$$

by substituting the appropriate eigenvalue  $\lambda_i$  and performing the matrix subtraction. Because  $\lambda_1$  is less illustrative than the other eigenvalues, let us work with  $\lambda_2$ . The result is

$$(\vec{P} - \lambda_2 \vec{I}) \vec{x}_2 = 0$$

$$\begin{bmatrix} -0.1924 & 0.5000 & 0.1000 \\ 0.1000 & -0.0924 & 0.4000 \\ 0.0100 & 0.0900 & 0.3076 \end{bmatrix} \begin{bmatrix} x_{12} \\ x_{22} \\ x_{32} \end{bmatrix} = 0$$

This represents a homogeneous linear system. Solution by elimination produces the relations,

$$\begin{aligned} x_{12} &= -6.4909 x_{32} \\ x_{22} &= -2.6971 x_{32} \end{aligned}$$

This result shows that to each eigenvalue there corresponds an infinite number of eigenvectors. We can obtain one of them by selecting  $x_{32} = 1$ . Then,

$$\vec{x}_2 = \begin{bmatrix} x_{12} \\ x_{22} \\ x_{32} \end{bmatrix} = \begin{bmatrix} -6.4909 \\ -2.6971 \\ 1.0000 \end{bmatrix}$$

We can normalize this eigenvector in the usual way by having its largest element be +1. This can be accomplished by dividing all the elements by a value equal to the first element. Then,

$$\vec{x}_2 = \begin{bmatrix} 1.0000 \\ 0.4152 \\ -0.1541 \end{bmatrix}$$

By a similar process, we can obtain the eigenvectors corresponding to the remaining eigenvalues:

$$\vec{x}_1 = \begin{bmatrix} 1.0000 \\ 1.0000 \\ 1.0000 \end{bmatrix} \quad \vec{x}_3 = \begin{bmatrix} 1.0000 \\ -0.3920 \\ 0.0365 \end{bmatrix}$$

The eigenvectors  $\vec{x}_i$  are called "left" eigenvectors. We can find the "right" eigenvectors  $\vec{y}_i$  by taking the transpose  $\vec{P}^T$  of the matrix  $\vec{P}$  and solving it, as above, for eigenvalues and eigenvectors. The eigenvalues of the transpose will be the same as those of the original matrix, but the eigenvectors will be different. Indeed, we find

$$\begin{aligned} \vec{y}_1 &= [0.0560 \quad 0.2360 \quad 1.0000] \\ \vec{y}_2 &= [-0.3078 \quad -0.6922 \quad 1.0000] \\ \vec{y}_3 &= [-0.4935 \quad 1.0000 \quad -0.5064] \end{aligned}$$

To obtain the powers of  $\vec{P}$ , we need the matrix products  $\vec{x}_i \vec{y}_i$  and  $\vec{y}_i \vec{x}_i$ . Simple matrix multiplication produces

$$\begin{aligned} \vec{x}_1 \vec{y}_1 &= \begin{bmatrix} 0.0560 & 0.2360 & 1.0000 \\ 0.0560 & 0.2360 & 1.0000 \\ 0.0560 & 0.2360 & 1.0000 \end{bmatrix} & \vec{y}_1 \vec{x}_1 &= 1.2920 \\ \vec{x}_2 \vec{y}_2 &= \begin{bmatrix} -0.3078 & -0.6922 & 1.0000 \\ -0.1279 & -0.2876 & 0.4155 \\ 0.0474 & 0.1066 & -0.1541 \end{bmatrix} & \vec{y}_2 \vec{x}_2 &= -0.7495 \\ \vec{x}_3 \vec{y}_3 &= \begin{bmatrix} -0.4935 & 1.0000 & -0.5065 \\ 0.1935 & -0.3920 & 0.1985 \\ -0.0180 & 0.0365 & -0.0185 \end{bmatrix} & \vec{y}_3 \vec{x}_3 &= -0.9041 \end{aligned}$$

We can obtain the 12th power of the matrix  $\vec{P}$  by computing the following:

$$\begin{aligned} \vec{P}^{12} &= \frac{1^{12}}{1.2920} \begin{bmatrix} 0.0560 & 0.2360 & 1.0000 \\ 0.0560 & 0.2360 & 1.0000 \\ 0.0560 & 0.2360 & 1.0000 \end{bmatrix} \\ &\quad - \frac{0.5924^{12}}{0.7495} \begin{bmatrix} -0.3078 & -0.6922 & 1.0000 \\ -0.1279 & -0.2876 & 0.4155 \\ 0.0474 & 0.1066 & -0.1541 \end{bmatrix} \\ &\quad - \frac{0.2076^{12}}{0.9041} \begin{bmatrix} -0.4935 & 1.0000 & -0.5065 \\ 0.1935 & -0.3920 & 0.1985 \\ -0.0180 & 0.0365 & -0.0185 \end{bmatrix} \end{aligned}$$

with the result that

$$\vec{P}^{12} = \begin{bmatrix} 0.0441 & 0.1844 & 0.7715 \\ 0.0437 & 0.1834 & 0.7730 \\ 0.0432 & 0.1824 & 0.7744 \end{bmatrix}$$

Note that this result is exactly the same as  $\bar{P}^{12}$  calculated earlier in this chapter by explicit matrix manipulation. To implement this eigenvalue matrix multiplication method requires storing either in core or on peripheral storage several  $m \times m$  matrices  $\bar{X}_i$ . As shown in section 13 of this chapter, it is almost never necessary to retain all  $m$  of the matrices, since the higher order forms contribute insignificantly to  $\bar{P}^n$  or  $\bar{B}^n$ .

#### 11. An Equivalent Markov Model Based on REEP.

We have seen that it is in principle possible to develop a prediction scheme in which weather changes are modeled as a discrete time Markov process and which is based on a classical Markov transition matrix  $\bar{P}$  whose powers  $\bar{P}^n$  are used to prepare forecasts for  $n$  time steps in the future. The skill of such a model will depend on the extent to which the weather behaves as required in the Markov process.

There is an important practical limitation on the application of a classical Markov transition matrix to real problems of prediction. The limitation is that the size of the transition matrix itself grows exponentially with the number of predictors/predictands and the resolution of each predictor/predictand. Consider, for example, a weather forecasting scheme involving the cloud ceiling in five classes  $z_1$  through  $z_5$ , visibility in five classes  $z_6$  through  $z_{10}$ , wind in nine classes  $z_{11}$  through  $z_{19}$ , and altimeter setting in ten classes  $z_{20}$  through  $z_{29}$ . Under these circumstances, the present or future weather is considered to be described by a 29-element binary vector (zeroes and ones). This is a conservative scheme; actual prediction models would typically encompass an order of magnitude greater number of binary or so called "dummy" variables. But let us continue with this example for illustrative purposes.

To apply the classical Markov approach, we must determine how many Markov states there are in the problem. Each combination among the 29 elements of the observation vector constitutes one such state, for example:

State Number	$z_1$	$z_2$	$z_3$	$z_4$	$z_5$	...	$z_p$	...	$z_{p-1}$	$z_p$
1	1	0	0	0	0		0		0	0
2	0	1	0	0	0		0		0	0
3	0	0	1	0	0		0		0	0
...	...	...	...	...	...		...		...	...

Each of the variables  $z_p$  has two possible outcomes, zero or one ("off" or "on"). The dummy variables in any one variable category (such as  $z_6$  through  $z_{10}$  for the visibility category) are mutually exclusive and exhaustive. Thus, within any category, one and only one  $z$  must be "on" at a time. Under these circumstances, considering the first variable category alone, there are five Markov states. For each of these, the second category produces five further states, for a total of  $5 \cdot 5 = 25$  Markov states. For each of these 25 states, the third category produces nine, giving a total of  $25 \cdot 9 = 225$  states. The final category of ten classes brings the total to  $225 \cdot 10 = 2,250$  Markov states! If  $n_c$  is the number of binary variables  $z_p$  in the  $c$ th category, and if there are  $C$  categories, then the number of Markov states  $M$  is given by

$$M = \prod_{c=1}^C n_c$$

For the easily imagined case of ten categories each containing ten dummy variable classes, the number of Markov states is  $1 \times 10^{10}$  or 10 billion! Clearly the pursuit of a Markov transition matrix of size (10 billion  $\cdot$  10 billion) is a hopeless exercise. An alternative is needed for all except the simplest forecasting problems.

A prediction method that yields probabilistic forecasts comparable to those of the classical Markov method but without the necessity of preparing a Markov transition matrix has been proposed by Miller (1968) and used in the forecasting experiment described in chapter 10 of this report. Because Miller's method is based on the Markov requirement for dependence of the current trial only on the immediately preceding trial, because the predictions turn out to be comparable, and because the proposed method also involves taking powers of a matrix, we will refer to Miller's technique as an equivalent Markov model. This method appears to retain much of the beauty and simplicity of the classical finite Markov chain while avoiding the overwhelming complexity occasioned by the  $n$ -dimensionality in real world observations of the state of physical systems.

Consider, as before, that the present and future weather can be described in terms of  $C$  classes of dummy variables  $z_p$ . The  $c$ th class contains  $n_c$  dummies, and

$$P = \sum_{c=1}^C n_c$$



the total number of dummies. For mathematical convenience, an extra dummy variable  $z_0$  is included in the observation and forecast vectors  $\vec{z}$ .<sup>\*</sup> Thus,

$$\vec{z} = [z_0 \ z_1 \ z_2 \ \dots \ z_p \ \dots \ z_{p-1} \ z_p]$$

Regression estimation of event probabilities (REEP, see chapter 5 of this report) makes it possible to estimate the likelihood of occurrence of condition  $z_p$  at time  $t$ , given the observed conditions at time  $t_0$ , i.e.,

$$\Pr(z_{p,t=1} | \vec{z}_{t_0})$$

by means of regression equation of the form,

$$\Pr(z_{p,t=1} | \vec{z}_{t_0}) = b_{p,0} z_{0,t_0} + b_{p,1} z_{1,t_0} + b_{p,2} z_{2,t_0} + \dots + b_{p,p} z_{p,t_0}$$

where the coefficients  $b_{p,p}$  are determined by a least squares technique and where the dummy variables  $z_p$  can be reduced in number by application of a method such as screening regression.

In the equivalent Markov approach, there must be as many predictands and prediction equations as there are predictors, including  $z_0$ , which is always unity both as a predictor and as a predictand. Thus, a system of multiple linear regression equations of the REEP variety emerges as follows:

$$\begin{aligned} \Pr(z_{0,t=1} | \vec{z}_{t_0}) &= b_{0,0} z_{0,t_0} + b_{0,1} z_{1,t_0} + b_{0,2} z_{2,t_0} + \dots + b_{0,p} z_{p,t_0} \\ \Pr(z_{1,t=1} | \vec{z}_{t_0}) &= b_{1,0} z_{0,t_0} + b_{1,1} z_{1,t_0} + b_{1,2} z_{2,t_0} + \dots + b_{1,p} z_{p,t_0} \\ \Pr(z_{2,t=1} | \vec{z}_{t_0}) &= b_{2,0} z_{0,t_0} + b_{2,1} z_{1,t_0} + b_{2,2} z_{2,t_0} + \dots + b_{2,p} z_{p,t_0} \\ &\dots \qquad \dots \qquad \dots \qquad \dots \qquad \dots \\ \Pr(z_{p,t=1} | \vec{z}_{t_0}) &= b_{p,0} z_{0,t_0} + b_{p,1} z_{1,t_0} + b_{p,2} z_{2,t_0} + \dots + b_{p,p} z_{p,t_0} \end{aligned}$$

In the example above involving 29 classes of ceiling, visibility, wind and altimeter setting,  $P = 29$ , and there are 30 equations in 30 terms.\*\* These equations form one system of equations, not four separate systems for the four variable categories. Binding a variety of predictors and predictands together in a single, jointly determined system gives the prediction scheme greater skill than would be attainable using separate systems of equations.

This system of REEP equations can conveniently be represented in matrix form as

$$\vec{R} = \vec{B} \vec{O}$$

where  $\vec{R}$  is the  $(P+1) \cdot 1$  column vector of probabilities of  $z_0, z_1, z_2 \dots z_p$  being "on" at time  $t = t_0 + \Delta t$ .  $\vec{O}$  is the similarly dimensioned column vector of observations  $z_0, z_1, z_2 \dots z_p$  at time  $t_0$ .  $\vec{B}$  is the  $(P+1) \cdot (P+1)$  matrix of REEP coefficients and represents a probability generating function:

<sup>\*</sup> $z_0$  is always unity in any observation or forecast.

<sup>\*\*</sup>The added term and equation is due to  $z_0$ .  $b_{0,0}$  must be unity, and all other  $b_0$  terms must be zero.

$$\vec{B} = \begin{bmatrix} b_{0,0} & b_{0,1} & b_{0,2} & \dots & b_{0,p} \\ b_{1,0} & b_{1,1} & b_{1,2} & \dots & b_{1,p} \\ b_{2,0} & b_{2,1} & b_{2,2} & \dots & b_{2,p} \\ \dots & \dots & \dots & \dots & \dots \\ b_{p,0} & b_{p,1} & b_{p,2} & \dots & b_{p,p} \end{bmatrix}$$

To obtain the forecast probabilities  $\vec{R}$  of the state of the system at time  $t = t_0 + \Delta t$ , we need only postmultiply the matrix  $\vec{B}$  with the observation vector  $\vec{o}$  describing the weather at time  $t_0$ . To obtain a forecast for a time  $t = t_0 + n \Delta t$ , which is  $n$  time steps in the future, one simply uses the  $n$ th power of  $\vec{B}$  as in the classical Markov technique:

$$\vec{R} = \vec{B}^n \vec{o}$$

where the powers can be obtained by the eigenvalue technique discussed above.

It should be noted that although powers of  $\vec{B}$  are used to obtain the forecast probabilities  $\vec{R}$  for several time steps in the future, the matrix  $\vec{B}$  is nevertheless not equal to the Markov transition matrix  $\vec{P}$ . It is rather the case that the forecast probabilities  $\vec{R}$  are comparable to those  $\vec{q} \vec{P}^n$  produced by the classical Markov method. For a prediction scheme in one category, the forecast probabilities are not just comparable; they are identical:

$$\vec{q} \vec{P}^n = \vec{B}^n \vec{o}$$

We have made the point that equivalent Markov forecasts are not in general identical to classical Markov forecasts, and we have used the term "comparable" to describe the agreement between the two methods. Why don't the methods agree exactly? The reason is that no 100 x 100 matrix of REEP coefficients  $\vec{B}$  can be expected to reproduce fully all the non-linear atmospheric state change relationships embodied in a 10 billion x 10 billion Markov transition matrix  $\vec{P}$ . The REEP equations in effect linearize the prediction scheme by neglecting the nonlinear Boolean combinations of dummy variables that constitute most of the Markov states. If every such Boolean combination were resurrected and made use of as a predictor in the equivalent Markov scheme, then the latter would produce forecasts in exact agreement with those of the classical Markov model. Of course, at that point the equivalent Markov scheme would be just as unwieldy in a practical forecast situation as the classical Markov technique it is intended to replace. In practice, since the equivalent Markov model is quite skillful in forecasting the weather (see chapter 10), it is apparent that linearizing the prediction scheme does not unacceptably impair the prognostic performance of the model.

One special case exists in which the predictions of the equivalent Markov model agree exactly with those of the classical technique. This is the case where the prediction scheme includes only one variable category (e.g., ceiling alone or visibility alone, with no "additional" predictors). In this case, each Markov state of the classical technique corresponds to exactly one binary variable of the equivalent Markov model and the classical Markov model.

The equivalent Markov model has a number of features of special interest that make it attractive from the point of view of practical prediction:

- When using the equivalent Markov method, it is unnecessary to prepare an explicit Markov transition probability matrix, which is often impracticably large in real world prediction problems.
- The probability generating function  $\vec{B}$ , when subjected to eigenvalue analysis, reveals the component of the forecast probability vector due to climatology and the components due to higher order, mean-departure influences. The contribution of climatology relative to the other terms is always readily obtainable in terms of ratios of powers of normalized eigenvalues. Convergence of the Markov process toward the steady climatological state is readily apparent through the eigenvalue analysis.
- The model can accommodate uncertainty in specification of the initial state. As is the case with the classical Markov method, one simply uses a probabilistic initial vector instead of a deterministic one when conditions are not known exactly. Thus, missing observations do not cripple the prediction scheme.

- Through the method of generalized statistical operators, discussed below, a single equivalent Markov model can be made applicable to large regions or even to the whole globe, making it unnecessary to prepare probability generating matrices  $B$  for every weather station. Network operators can also be used if necessary.
- Non-linear Boolean combinations of predictors/predictands as well as time lags can be incorporated in the model. The time lag feature makes it possible in part to circumvent the classical Markov model's independence of the outcome of trials earlier than that immediately preceding.
- With wise choice of predictors and predictands, the equivalent Markov model can be made to accommodate synoptic data such as pilot reports, radar weather observations and satellite information. The model can be run at any time a forecast is needed.
- The skill of the prediction scheme can be improved by incorporation of predictors beyond the minimum set dictated by the requirement that each predictand must also serve as a predictor. Naturally, when "additional" predictors are made use of, corresponding "additional" predictands appear. In practice, this is handled either by disregarding the superfluous predictions or by abbreviating the forecast algorithm (matrix multiplication) such that the unneeded predictions are simply not made.
- The model is well suited for the use of a special class of predictors referred to as model output statistics (MOS). Typically, MOS predictors are produced by numerical weather prediction models and represent forecast values of various atmospheric parameters applicable at some future time. Examples useful in visibility forecasting, to consider one application, include vertical motion at 850 mb and winds at the top of the Ekman layer. Under the "imperfect prog" or MOS-approach to forecasting, statistical relations are developed between these MOS predictors and the occurrence of the weather elements being forecast, such as ceiling and visibility. The equivalent Markov model can handle both MOS predictors (imperfect prog approach) and observed predictors (perfect prog approach) in the same model equations. In the MOS predictors selected are appropriate to the prediction task at hand, it can be expected that the skill of the statistical scheme will ride along on the skill of the dynamical model.

## 12. Equivalent Markov and Classical Markov Models: A Comparative Example.

It is not obvious that the equivalent Markov model presented above is in fact equivalent to the classical finite Markov chain model given earlier. Rather than to prove the equivalency mathematically, let us demonstrate it by means of an example. While less rigorous than a proof, the example will highlight certain computational methods useful in application of the equivalent Markov technique.

Consider the problem of forecasting the cloud ceiling at Kelly AFB, Texas. For simplicity, we will pose this problem in terms of one predictor/predictand category, the ceiling itself. We will subdivide the one category into three classes,  $z_1$ ,  $z_2$  and  $z_3$ , as follows:

Table 1

### DUMMY VARIABLES AND MARKOV STATES FOR CEILING HEIGHT PROBLEM

Dummy Variable	Definition	Markov State
$z_1$	$0 \text{ ft} \leq \text{CIG} < 3,000 \text{ ft}$	1
$z_2$	$3,000 \text{ ft} \leq \text{CIG} < 15,000 \text{ ft}$	2
$z_3$	$15,000 \text{ ft} \leq \text{CIG}$	3

Because this problem has only one category of predictor/predictand, the Markov states exactly parallel the dummy variables. This convenient property does not hold true for problems with more than one predictor/predictand, and the lack of correspondence between dummies and states makes more complex any comparison between equivalent Markov and classical Markov techniques. The classical Markov method forecasts, for example, the probability of occurrence of Markov state 12, which is a particular combination of dummy variables being "on," say  $z_3$ ,  $z_4$  and  $z_{31}$ . To get the probability of one such dummy, say  $z_4$ , being on, we would have to add the classical Markov probability forecasts for any Markov states in which  $z_4$  is categorized as "on." Only then would we have a forecast probability of  $z_4$  for comparison with the equivalent Markov forecast. It is to avoid this final step of adding the classical Markov predictions that we have limited the number of variable categories in this example to one.



To permit the reader to verify these results by hand, the data base for this example has been limited to 48 hours of actual ceiling observations for Kelly AFB, TX, during the period 11-12 April 1950. These are reproduced in Table 2. In practice, such a data base is much too small. In general, periods of record four orders of magnitude longer than this are used to insure robustness of the prediction scheme when applied to independent data.

Table 2

48 HOURS OF OBSERVATIONS OF CLOUD CEILING AT KELLY AFB, TX

11-12 April 1950

<u>Date/Time (L)</u>	<u>Ceiling (ft)</u>	<u>Markov State</u>
11/00	800	1
01	800	1
02	800	1
03	2,500	1
04	2,600	1
05	1,700	1
06	2,300	1
07	3,000	2
08	3,000	2
09	4,500	2
10	25,000	3
11	25,000	3
12	25,000	3
13	25,000	3
14	None	3
15	None	3
16	None	3
17	None	3
18	None	3
19	25,000	3
20	25,000	3
21	5,000	2
22	4,500	2
23	4,000	2
12/00	10,000	2
01	11,000	2
02	900	1
03	800	1
04	7,000	2
05	1,000	1
06	7,000	2
07	500	1
08	600	1
09	7,000	2
10	1,900	1
11	1,900	1
12	25,000	3
13	25,000	3
14	25,000	3
15	25,000	3
16	25,000	3
17	25,000	3
18	25,000	3
19	25,000	3
20	25,000	3
21	25,000	3
22	25,000	3
23	1,900	1
13/00	1,600	1

Explicit counting of Markov state transitions in the data of Table 2 gives rise to the classical Markov transition probability matrix  $\bar{P}$  shown in Table 3. Note the dominance of persistence along the main diagonal. Note also for the data base selected how unlikely it is that bad weather will improve (see row 1). In this case, probably because of the limited data base used, the Markov model will strongly predict deteriorating weather, given that conditions are already marginal (see row 2). The greater dominance of persistence in good weather regimes is seen in row 3.

Table 3

MARKOV TRANSITION MATRIX\*  
FORMED BY EXPLICIT COUNTING OF STATE CHANGES

		Final Markov State			
		<u>1</u>	<u>2</u>	<u>3</u>	
Initial Markov State	<u>1</u>	0.6667 (10)	0.2667 (4)	0.0667 (1)	1.0000 (15)
	<u>2</u>	0.3636 (4)	0.5455 (6)	0.0909 (1)	1.0000 (11)
	<u>3</u>	0.0455 (1)	0.0455 (1)	0.9091 (20)	1.0000 (22)

It is apparent that we can make a 1-hour weather forecast for time  $t = t_0 + \Delta t$  directly from the transition probability matrix  $\bar{P}$  in Table 3. The three possible forecasts, generated from the three possible initial states, are shown in the first part of Table 4 below. It is likewise possible to make a 6-hour forecast for time  $t = t_0 + 6 \Delta t$  from the sixth power of  $\bar{P}$ , where

$$\bar{P}^6 = \begin{bmatrix} 0.4082 & 0.2924 & 0.2994 \\ 0.3987 & 0.2871 & 0.3142 \\ 0.2041 & 0.1571 & 0.6388 \end{bmatrix}$$

The three possible 6-hour forecasts are also shown in Table 4:

Table 4

		Initial State			Probability of Final State		
Time		<u>1</u>	<u>2</u>	<u>3</u>	<u>1</u>	<u>2</u>	<u>3</u>
1-hour	1	0	0		0.6667	0.2667	0.0667
	0	1	0		0.3636	0.5455	0.0909
	0	0	1		0.0455	0.0455	0.9091
6-hour	1	0	0		0.4082	0.2924	0.2994
	0	1	0		0.3987	0.2871	0.3142
	0	0	1		0.2041	0.1571	0.6388

The question is, can these probabilistic forecasts be reproduced by the equivalent Markov model presented above? To see whether this is possible, let us first transform our data into binary, or so called "dummy" variables, including a  $z_0$  variable that is always on. Note that the final state for the  $n$ th observation also serves as the initial state for the  $(n+1)$ th observation. The data expressed in terms of dummy variables are shown in Table 5.

\*Transition probabilities are shown as fractional values. Counts are shown as integers in parentheses.

Table 5

MARKOV ANALYSIS IN TERMS OF DUMMY VARIABLES  
SUITABLE FOR USE IN MULTIPLE LINEAR REGRESSION

Observation	Initial State ( $t_0$ )				Final State ( $t$ )			
	$z_{0,t_0}$	$z_{1,t_0}$	$z_{2,t_0}$	$z_{3,t_0}$	$z_{0,t}$	$z_{1,t}$	$z_{2,t}$	$z_{3,t}$
1	1	1	0	0	1	1	0	0
2	1	1	0	0	1	1	0	0
3	1	1	0	0	1	1	0	0
4	1	1	0	0	1	1	0	0
5	1	1	0	0	1	1	0	0
6	1	1	0	0	1	1	0	0
7	1	1	0	0	1	0	1	0
8	1	0	1	0	1	0	1	0
9	1	0	1	0	1	0	1	0
10	1	0	1	0	1	0	0	1
11	1	0	0	1	1	0	0	1
12	1	0	0	1	1	0	0	1
13	1	0	0	1	1	0	0	1
14	1	0	0	1	1	0	0	1
15	1	0	0	1	1	0	0	1
16	1	0	0	1	1	0	0	1
17	1	0	0	1	1	0	0	1
18	1	0	0	1	1	0	0	1
19	1	0	0	1	1	0	0	1
20	1	0	0	1	1	0	0	1
21	1	0	0	1	1	0	1	0
22	1	0	1	0	1	0	1	0
23	1	0	1	0	1	0	1	0
24	1	0	1	0	1	0	1	0
25	1	0	1	0	1	0	1	0
26	1	0	1	0	1	1	0	0
27	1	1	0	0	1	1	0	0
28	1	1	0	0	1	0	1	0
29	1	0	1	0	1	1	0	0
30	1	1	0	0	1	0	1	0
31	1	0	1	0	1	1	0	0
32	1	1	0	0	1	1	0	0
33	1	1	0	0	1	0	1	0
34	1	0	1	0	1	1	0	0
35	1	1	0	0	1	1	0	0
36	1	1	0	0	1	0	0	1
37	1	0	0	1	1	0	0	1
38	1	0	0	1	1	0	0	1
39	1	0	0	1	1	0	0	1
40	1	0	0	1	1	0	0	1
41	1	0	0	1	1	0	0	1
42	1	0	0	1	1	0	0	1
43	1	0	0	1	1	0	0	1
44	1	0	0	1	1	0	0	1
45	1	0	0	1	1	0	0	1
46	1	0	0	1	1	0	0	1
47	1	0	0	1	1	1	0	0
48	1	1	0	0	1	1	0	0

From Table 5, the binary nature of the dummy variables makes it exceedingly simple to obtain a sum of the squares and cross products matrix, such as that below. Such matrices are also called



SSCP matrices. Note that this matrix is symmetrical. It is economical to retain only the upper or lower triangular elements even though this may seem to make the programming more complex.

$$S = \begin{bmatrix} \Sigma z_{0,t_0}^2 & \Sigma z_{1,t_0} & \Sigma z_{2,t_0} & \Sigma z_{3,t_0} & \Sigma y_t \\ \Sigma z_{1,t_0} & \Sigma z_{1,t_0}^2 & \Sigma z_{1,t_0} z_{2,t_0} & \Sigma z_{1,t_0} z_{3,t_0} & \Sigma y_t z_{1,t_0} \\ \Sigma z_{2,t_0} & \Sigma z_{1,t_0} z_{2,t_0} & \Sigma z_{2,t_0}^2 & \Sigma z_{2,t_0} z_{3,t_0} & \Sigma y_t z_{2,t_0} \\ \Sigma z_{3,t_0} & \Sigma z_{1,t_0} z_{3,t_0} & \Sigma z_{2,t_0} z_{3,t_0} & \Sigma z_{3,t_0}^2 & \Sigma y_t z_{3,t_0} \\ \Sigma y_t & \Sigma z_{1,t_0} y_t & \Sigma z_{2,t_0} y_t & \Sigma z_{3,t_0} y_t & \Sigma y_t^2 \end{bmatrix}$$

In this notation, the second subscript on the z's, being  $t_0$ , emphasizes the point that this is predictor information, available at time  $t_0$ . The variable y represents the predictand, valid at time t. A y is used instead of z because the y stands for any of the three predictand z's:

$$y = z_1 \text{ or } z_2 \text{ or } z_3$$

There accordingly exist three y-rows for the SSCP matrix  $\vec{S}$ . Any one matrix  $\vec{S}$  can accommodate only one y-row, so in fact there are three SSCP matrices,  $\vec{S}_1$ ,  $\vec{S}_2$  and  $\vec{S}_3$ , each having the same first four rows and a different y-row. Each matrix  $\vec{S}$  will serve indirectly as the basis for development of one REEP equation. In computational practice, the duplication among the three matrices  $\vec{S}$  makes it unnecessary to carry all three of them, but we show them as distinct in Table 6 for illustrative purposes.

Table 6  
SSCP MATRICES  $\vec{S}$  AND LEFT-OUT MATRICES  $\vec{L}$   
FOR EQUIVALENT MARKOV PROBLEM  
(Prepared from Data in Table 5)

$$\begin{aligned} \vec{S}_1 &= \begin{bmatrix} 48 & 15 & 11 & 22 & 15 \\ 15 & 15 & 0 & 0 & 10 \\ 11 & 0 & 11 & 0 & 4 \\ 22 & 0 & 0 & 22 & 1 \\ 15 & 10 & 4 & 1 & 15 \end{bmatrix} & \vec{L}_1 &= \begin{bmatrix} 48 & 15 & 11 & 15 \\ 15 & 15 & 0 & 10 \\ 11 & 0 & 11 & 4 \\ 15 & 10 & 4 & 15 \end{bmatrix} \\ \vec{S}_2 &= \begin{bmatrix} 48 & 15 & 11 & 22 & 11 \\ 15 & 15 & 0 & 0 & 4 \\ 11 & 0 & 11 & 0 & 6 \\ 22 & 0 & 0 & 22 & 1 \\ 11 & 4 & 6 & 1 & 11 \end{bmatrix} & \vec{L}_2 &= \begin{bmatrix} 48 & 15 & 11 & 11 \\ 15 & 15 & 0 & 4 \\ 11 & 0 & 11 & 6 \\ 11 & 4 & 6 & 11 \end{bmatrix} \\ \vec{S}_3 &= \begin{bmatrix} 48 & 15 & 11 & 22 & 22 \\ 15 & 15 & 0 & 0 & 1 \\ 11 & 0 & 11 & 0 & 1 \\ 22 & 0 & 0 & 22 & 20 \\ 22 & 1 & 1 & 20 & 22 \end{bmatrix} & \vec{L}_3 &= \begin{bmatrix} 48 & 15 & 11 & 22 \\ 15 & 15 & 0 & 1 \\ 11 & 0 & 11 & 1 \\ 22 & 1 & 1 & 22 \end{bmatrix} \end{aligned}$$

The REEP equations cannot be derived directly from the SSCP matrices  $\vec{S}$  because the use of mutually exclusive and exhaustive dummy variables gives rise to redundant rows and columns in such matrices. For example, in the present case, all the information about the state of variable  $z_3$  is given by the states of  $z_1$  and  $z_2$ . We must remove from each matrix  $\vec{S}$  one redundant row and column for each category of variable. In the present example, there is only one variable category, the ceiling; thus, we must delete one row and one column. Arbitrarily, we select the row and column corresponding to  $z_3$  for deletion. The reduced matrices  $\vec{L}$  in "left-out variable" form are also shown in Table 6.

Statistics texts such as Tatsuoka (1971) present the matrix formulation of multiple linear regression problems. In the present case, we require only two regression equations in left-out variable form:

$$\Pr(z_{1,t=1}|\vec{z}_{t_0}) = a_{1,0}z_{0,t_0} + a_{1,1}z_{1,t_0} + a_{1,2}z_{2,t_0}$$

$$\Pr(z_{2,t=1}|\vec{z}_{t_0}) = a_{2,0}z_{0,t_0} + a_{2,1}z_{1,t_0} + a_{2,2}z_{2,t_0}$$

where the regression equations are given by elements of the Crout auxiliary matrix  $\vec{D}$

$$\vec{D} = \begin{bmatrix} d_{1,1} & d_{1,2} & d_{1,3} & d_{1,4} \\ d_{2,1} & d_{2,2} & d_{2,3} & d_{2,4} \\ d_{3,1} & d_{3,2} & d_{3,3} & d_{3,4} \\ d_{4,1} & d_{4,2} & d_{4,3} & d_{4,4} \end{bmatrix}$$

whose creation is discussed in chapter 2 above. In this case,

$$d_{11} = 1_{11}$$

$$d_{21} = 1_{21}$$

$$d_{31} = 1_{31}$$

$$d_{41} = 1_{41}$$

$$d_{12} = d_{21} / d_{11}$$

$$d_{13} = d_{31} / d_{11}$$

$$d_{14} = d_{41} / d_{11}$$

$$d_{22} = 1_{22} - d_{12}d_{21}$$

$$d_{32} = 1_{32} - d_{12}d_{31}$$

$$d_{42} = 1_{42} - d_{12}d_{41}$$

$$d_{23} = d_{32} / d_{22}$$

$$d_{24} = d_{42} / d_{22}$$

$$d_{33} = 1_{33} - d_{13}d_{31} - d_{23}d_{32}$$

$$d_{43} = 1_{43} - d_{13}d_{41} - d_{23}d_{42}$$

$$d_{34} = d_{43} / d_{33}$$

$$d_{44} = 1_{44} - d_{41}d_{14} - d_{42}d_{24} - d_{43}d_{34}$$

and the regression coefficients are

$$a_{1,2} = d_{3,4}$$

$$a_{1,1} = d_{2,4} - d_{2,3}a_{1,2}$$

$$a_{1,0} = d_{1,4} - d_{1,3}a_{1,2} - d_{1,2}a_{1,1}$$

We shall show the computation of the regression coefficients  $a_{ij}$  for the first "left-out" regression equation and allow the reader to supply corresponding detail for the other two equations. The Crout auxiliary matrix  $\vec{D}_1$  corresponding to the left-out SSCP matrix  $\vec{L}_1$  is

$$\vec{D}_1 = \begin{bmatrix} 48.0000 & 0.3125 & 0.2292 & 0.3125 \\ 15.0000 & 10.3125 & -0.3333 & 0.5152 \\ 11.0000 & -3.4375 & 7.3333 & 0.3182 \\ 15.0000 & 5.3125 & 2.3333 & 6.8333 \end{bmatrix}$$

The left-out regression coefficients are then

$$a_{1,0} = 0.04545 \quad a_{1,1} = 0.6212 \quad a_{1,2} = 0.3182$$

Once the other regression coefficients have been computed, a left-out matrix of coefficients  $\vec{A}$  can be formed as follows:

$$\vec{A} = \begin{bmatrix} a_{0,0} & a_{0,1} & a_{0,2} \\ a_{1,0} & a_{1,1} & a_{1,2} \\ a_{2,0} & a_{2,1} & a_{2,2} \end{bmatrix} = \begin{bmatrix} 1.0000 & 0.0000 & 0.0000 \\ 0.0455 & 0.6212 & 0.3182 \\ 0.0455 & 0.2212 & 0.5000 \end{bmatrix}$$

Note that the first row is supplied to accommodate the trivial prediction equation,

$$\Pr(z_{0,t=1} | \vec{z}_{t_0}) = a_{0,0}z_{0,t_0} + a_{0,1}z_{1,t_0} + a_{0,2}z_{2,t_0}$$

This first, or "zero" row must always consist of a one in the first element and zeroes in the remaining elements because  $z_{0,t}$  and  $z_{0,t_0}$  are both always one.

Forecasts in the form

$$\vec{Pr} = \vec{A} \vec{z}$$

or

$$\begin{bmatrix} \Pr(z_{0,t=1} | \vec{z}_{t_0}) \\ \Pr(z_{1,t=1} | \vec{z}_{t_0}) \\ \Pr(z_{2,t=1} | \vec{z}_{t_0}) \end{bmatrix} = \begin{bmatrix} a_{0,0} & a_{0,1} & a_{0,2} \\ a_{1,0} & a_{1,1} & a_{1,2} \\ a_{2,0} & a_{2,1} & a_{2,2} \end{bmatrix} \begin{bmatrix} z_{0,t_0} \\ z_{1,t_0} \\ z_{2,t_0} \end{bmatrix}$$



can be made from these equations as they stand. For example, for the initial conditions

$$z_{0,t_0} = 1 \quad z_{1,t_0} = 0 \quad z_{2,t_0} = 1$$

the left-out form of the equation predicts

$$\Pr(z_{0,t}) = 1.0000 + 0.0000 + 0.0000 = 1.0000$$

$$\Pr(z_{1,t}) = 0.0455 + 0.0000 + 0.3182 = 0.3636$$

$$\Pr(z_{2,t}) = 0.0455 + 0.0000 + 0.5000 = 0.5455$$

which is the same as the second line of Table 4. We can obtain  $\Pr(z_{3,t})$  by subtraction:

$$\Pr(z_3) = 1 - \Pr(z_1) - \Pr(z_2)$$

This "left-out" form of the REEP equations is inconvenient, in that it does not directly produce a forecast for the left-out dummy variables except by subtraction. Moreover, it is not amenable to subsequent eigenvalue analysis. Ordinarily, therefore, one subjects the left-out matrix  $\vec{A}$  to a procedure known as PLODITE (putting the left-out dummies in the equation). The PLODITE algorithm produces a matrix of regression coefficients  $\vec{B}$  that includes a row and column corresponding to each left-out dummy variable.

A general PLODITE algorithm in the form of the FORTRAN program PLDT is provided in appendix A. We can describe that algorithm as it applies to the problem of treating our 3 x 3 matrix  $\vec{A}$ .

The first step is to identify the left-out variables and the rows and columns of  $\vec{B}$  corresponding to them. In the present case,  $z_3$  has been left out, so we must supply a new column  $b_{i,3}$  and row  $b_{3,j}$  in the matrix  $\vec{B}$  of REEP coefficients:

$$\vec{B} = \begin{array}{c} \begin{array}{c} \downarrow \\ \begin{bmatrix} b_{0,0} & b_{0,1} & b_{0,2} & b_{0,3} \\ b_{1,0} & b_{1,1} & b_{1,2} & b_{1,3} \\ b_{2,0} & b_{2,1} & b_{2,2} & b_{2,3} \\ \rightarrow \begin{bmatrix} b_{3,0} & b_{3,1} & b_{3,2} & b_{3,3} \end{bmatrix} \end{array} \end{array} \end{array}$$

The "zero row" of  $\vec{B}$  is trivial; it contains a one in the first element and zeroes in the others. The next two rows use the algorithm,

$$b_{i,3} = -a_{i,1}\bar{z}_{1,t_0} - a_{i,2}\bar{z}_{2,t_0} \quad \text{for } i > 0$$

$$= -a_{i,1} \frac{l_{1,2}}{l_{1,1}} - a_{i,2} \frac{l_{1,3}}{l_{1,1}}$$

$$= Q$$

$$b_{i,2} = a_{i,2} + Q \quad \text{for } i > 0$$

$$b_{i,1} = a_{i,1} + Q \quad \text{for } i > 0$$

$$b_{i,0} = \bar{z}_{i,t} \quad \text{for } i > 0$$

where  $\bar{z}_{i,t}$  is the mean of the  $i$ th predictand and can be obtained from the first element of the

AD-A050 035

AIR WEATHER SERVICE SCOTT AFB ILL  
SELECTED TOPICS IN STATISTICAL METEOROLOGY. (U)  
JUL 77 R G MILLER, B D ALTENHOF, J N FULFORD  
AWS-TR-77-273

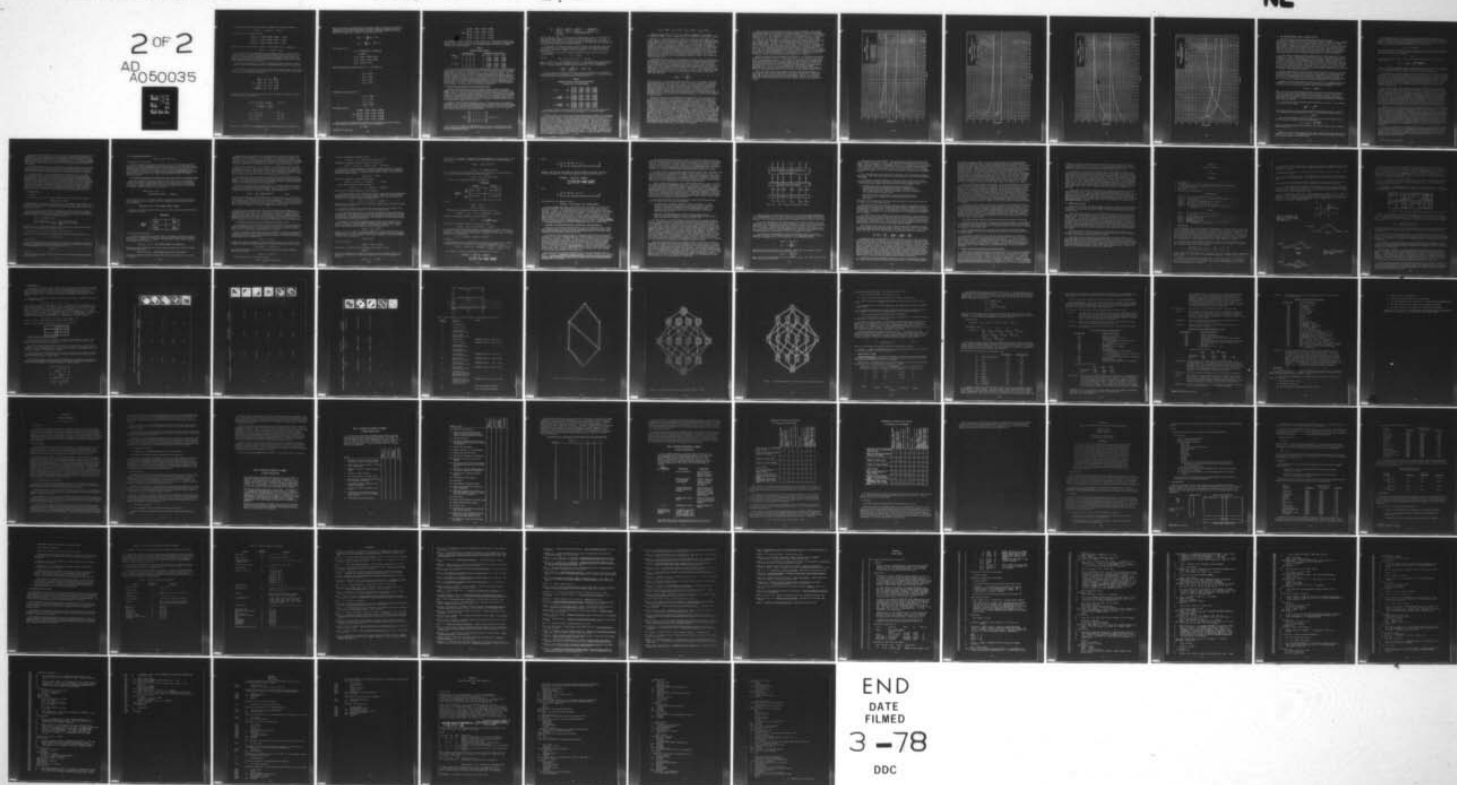
F/G 4/2

UNCLASSIFIED

NL

2 OF 2

AD  
A050035



END  
DATE  
FILMED

3-78

DDC

can be made from these equations as they stand. For example, for the initial conditions

$$z_{0,t_0} = 1 \quad z_{1,t_0} = 0 \quad z_{2,t_0} = 1$$

the left-out form of the equation predicts

$$\Pr(z_{0,t}) = 1.0000 + 0.0000 + 0.0000 = 1.0000$$

$$\Pr(z_{1,t}) = 0.0455 + 0.0000 + 0.3182 = 0.3636$$

$$\Pr(z_{2,t}) = 0.0455 + 0.0000 + 0.5000 = 0.5455$$

which is the same as the second line of Table 4. We can obtain  $\Pr(z_{3,t})$  by subtraction:

$$\Pr(z_3) = 1 - \Pr(z_1) - \Pr(z_2)$$

This "left-out" form of the REEP equations is inconvenient, in that it does not directly produce a forecast for the left-out dummy variables except by subtraction. Moreover, it is not amenable to subsequent eigenvalue analysis. Ordinarily, therefore, one subjects the left-out matrix  $\bar{A}$  to a procedure known as PLODITE (putting the left-out dummies in the equation). The PLODITE algorithm produces a matrix of regression coefficients  $\bar{B}$  that includes a row and column corresponding to each left-out dummy variable.

A general PLODITE algorithm in the form of the FORTRAN program PLDT is provided in appendix A. We can describe that algorithm as it applies to the problem of treating our 3 x 3 matrix  $\bar{A}$ .

The first step is to identify the left-out variables and the rows and columns of  $\bar{B}$  corresponding to them. In the present case,  $z_3$  has been left out, so we must supply a new column  $b_{i,3}$  and row  $b_{3,j}$  in the matrix  $\bar{B}$  of REEP coefficients:

$$\bar{B} = \begin{matrix} & & & \downarrow \\ & \begin{bmatrix} b_{0,0} & b_{0,1} & b_{0,2} & b_{0,3} \\ b_{1,0} & b_{1,1} & b_{1,2} & b_{1,3} \\ b_{2,0} & b_{2,1} & b_{2,2} & b_{2,3} \\ \rightarrow b_{3,0} & b_{3,1} & b_{3,2} & b_{3,3} \end{bmatrix} \end{matrix}$$

The "zero row" of  $\bar{B}$  is trivial; it contains a one in the first element and zeroes in the others. The next two rows use the algorithm,

$$\begin{aligned} b_{i,3} &= -a_{i,1}\bar{z}_{1,t_0} - a_{i,2}\bar{z}_{2,t_0} && \text{for } i > 0 \\ &= -a_{i,1} \frac{l_{1,2}}{l_{1,1}} - a_{i,2} \frac{l_{1,3}}{l_{1,1}} \\ &= Q \end{aligned}$$

$$b_{i,2} = a_{i,2} + Q \quad \text{for } i > 0$$

$$b_{i,1} = a_{i,1} + Q \quad \text{for } i > 0$$

$$b_{i,0} = \bar{z}_{i,t} \quad \text{for } i > 0$$

where  $\bar{z}_{i,t}$  is the mean of the  $i$ th predictand and can be obtained from the first element of the



final row of the left-out SSCP matrix  $\bar{L}_1$  by dividing the element by the number of observations  $N$ . The final row of  $\bar{B}$  in each variable category is obtained by summing the corresponding column elements in the rows above it, using the rule that the  $b_{1,0}$  must add columnwise to one within each variable category, while the other  $b_{1,j}$  must add to zero. Thus

$$b_{3,0} = 1 - \sum_{i=1}^2 b_{i,0} = \bar{z}_{3,t}$$

$$b_{3,j} = -\sum_{i=1}^2 b_{i,j} \quad \text{for } j > 0$$

For the first row  $i = 1$ ,

$$\begin{aligned} b_{1,3} &= -0.6212(15/48) - 0.3182(11/48) \\ &= -0.2670 = 0 \end{aligned}$$

$$b_{1,2} = 0.3182 - 0.2670 = 0.0511$$

$$b_{1,1} = 0.6212 - 0.2670 = 0.3542$$

$$b_{1,0} = (15/48) = 0.3125$$

When the same algorithm is applied to the row  $i = 2$ , the results are

$$b_{2,3} = -0.1837$$

$$b_{2,2} = 0.3163$$

$$b_{2,1} = 0.0375$$

$$b_{2,0} = 0.2292$$

By summation we obtain the row  $i = 3$ :

$$b_{3,3} = 0.4508$$

$$b_{3,2} = -0.3674$$

$$b_{3,1} = -0.3917$$

$$b_{3,0} = 0.4583$$

The complete matrix is

$$\bar{B} = \begin{bmatrix} 1.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.3125 & 0.3542 & 0.0511 & -0.2670 \\ 0.2292 & 0.0375 & 0.3163 & -0.1837 \\ 0.4583 & -0.3917 & -0.3674 & 0.4508 \end{bmatrix}$$

This is the matrix of REEP coefficients  $\bar{B}$  used in the equivalent Markov model to produce a forecast probability vector  $\bar{R}$  from the vector of initial conditions  $\bar{O}$ :

$$\bar{R} = \bar{B}^6 \bar{O}$$

Using  $\bar{B}$  and its sixth power,

$$\vec{B}^6 = \begin{bmatrix} 1.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.3125 & 0.0957 & 0.0862 & -0.1084 \\ 0.2292 & 0.0632 & 0.0579 & -0.0721 \\ 0.4583 & -0.1590 & -0.1441 & 0.1804 \end{bmatrix}$$

we can prepare forecasts for 1 hour ( $n = 1$ ) and 6 hours ( $n = 6$ ). These equivalent Markov forecasts are shown in Table 7, which is in the same form as Table 4 containing the classical Markov forecasts. Note that the results are identical, demonstrating the equivalence of the REEP technique and the classical method.

Table 7

EQUIVALENT MARKOV FORECASTS

Time	Initial State				Probability of Final State			
	0	1	2	3	0	1	2	3
1 - hour	1	1	0	0	1	0.6667	0.2667	0.0667
	1	0	1	0	1	0.3636	0.5455	0.0909
	1	0	0	1	1	0.0455	0.0455	0.9091
6 - hour	1	1	0	0	1	0.4082	0.2924	0.2944
	1	0	1	0	1	0.3987	0.2871	0.3142
	1	0	0	1	1	0.2041	0.1571	0.6388

In this section, we have demonstrated how an equivalent Markov probability generating function in the form of a REEP matrix of regression coefficients  $\vec{B}$  can be created from historical weather data and applied to a practical forecasting problem. A more extensive application of this technique is given in chapter 10. Our analysis is not complete at this point, however. The probability generating function  $\vec{B}$  is rich in information regarding how and under what conditions the weather changes, the degree to which it is predictable for different forecast periods, how strongly climatology contributes to the forecast probabilities for various forecast lengths, and the rate at which these forecast probabilities decay toward the vector of climatology. A part of this information can be elicited from  $\vec{B}$  by eigenvalue analysis of the type discussed above for obtaining powers of a matrix. In the following section, we will analyze the matrix  $\vec{B}$  developed from this example. Our analysis will be by the method of eigenvalues.

### 13. Hierarchical Matrix Analysis of the Equivalent Markov Model.

The Markov chains encountered in practical weather forecasting problems are completely ergodic in the sense that their limiting state probability distributions are independent of the initial conditions. In section 8 above, dealing with the classical Markov model, we saw this as a result of the fact that Markov transition matrices for weather forecasting problems are regular, bringing about a stationary distribution of the Markov chain after sufficiently many steps  $n$ . The forecast probabilities produced by the Markov model will, with increasing length of the forecast period, converge toward the steady state vector of a priori climatological probabilities as the system gradually "forgets" its initial state.

Convergence toward the steady climatological state, an attractive feature of the classical Markov model, is also seen in the equivalent Markov model. Although the matrix  $\vec{B}$  of REEP coefficients is not regular, nevertheless increasing powers of  $\vec{B}$  do converge to the zero-order hierarchical form  $\vec{C}_0$  having as its first column the climatological expectation of each of the predictor dummy variables and having zero in all other positions:

$$\vec{C}_0 = \begin{bmatrix} \bar{z}_0 & 0 & 0 & 0 & \dots & 0 \\ \bar{z}_1 & 0 & 0 & 0 & \dots & 0 \\ \bar{z}_2 & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \bar{z}_p & 0 & 0 & 0 & \dots & 0 \end{bmatrix} \quad \text{for } P = m - 1$$

As it turns out, the matrix of climatology  $\vec{C}_0$  and all the other hierarchical forms, which represent mean-departure terms, are given by the eigenvalue method for obtaining powers of the probability generating function  $\vec{B}$ :



$$\hat{B}^n = \underbrace{\frac{\lambda^n}{\bar{y}_0 \bar{x}_0} \bar{C}_0}_{\text{Mean Term "Climatology"}} + \underbrace{\frac{\lambda^n}{\bar{y}_1 \bar{x}_1} \bar{C}_1 + \frac{\lambda^n}{\bar{y}_2 \bar{x}_2} \bar{C}_2 + \dots + \frac{\lambda_{m-1}^n}{\bar{y}_{m-1} \bar{x}_{m-1}} \bar{C}_{m-1}}_{\text{Mean-Departure Terms}}$$

where  $\bar{C}_0 = \bar{x}_0 \bar{y}_0$ ,  $\bar{C}_1 = \bar{x}_1 \bar{y}_1$ , etc., and where  $m$  is the dimensionality of the  $m \times m$  matrix  $\bar{B}$ . The total number of dummy variables  $P$  is one less than  $m$  due to the presence of  $z_0$ .  $\lambda_i$  is the  $i$ th eigenvalue,  $\bar{x}_i$  its associated "left" eigenvector, and  $\bar{y}_i$  the associated "right" eigenvector. The  $\bar{C}_i$  are the hierarchical matrices of order  $i$  and have the same dimensions as the  $m \times m$  matrix  $B$ . Counting  $\bar{C}_0$ , there are formally  $m$  hierarchical forms  $\bar{C}_i$ .

The algorithm is the same as that given in section 10 above, except the index  $i$  has in this case been started at zero to emphasize the special role of  $\bar{C}_0$  as a zero-order term, being the climatological expectation or a priori, unconditional probability of occurrence of each of the dummy variables.

The multipliers of each of the hierarchical forms  $\bar{C}_i$  are scalars, which we can consider weighting factors:

$$w_i = \lambda_i^n / \bar{y}_i \bar{x}_i$$

Therefore,

$$\hat{B}^n = w_0 \bar{C}_0 + w_1 \bar{C}_1 + w_2 \bar{C}_2 + \dots + w_{m-1} \bar{C}_{m-1}$$

where  $w_i$  is always unity. The weighting factors  $w_i$  for  $i > 0$  become small with increasing forecast period (larger  $n$ ). At sufficiently large  $n$ , therefore, the contribution from the higher order terms  $i > 0$  becomes negligible, and  $\hat{B}^n$  converges to climatology, i.e.,  $\bar{C}_0$ . Mathematically,

$$\lim_{n \rightarrow \infty} \hat{B}^n = \lim_{n \rightarrow \infty} \sum_{i=0}^{m-1} w_i \bar{C}_i = w_0 \bar{C}_0 = \bar{C}_0$$

We can see this in the Kelly AFB cloud ceiling example by subjecting the  $4 \times 4$  probability generating function  $\bar{B}$  for that problem to hierarchical eigenvalue analysis. The results, in terms of hierarchical matrices and weighting factors, are shown in Table 8.

**Table 8**  
**EIGENVALUE ANALYSIS OF REEP COEFFICIENT MATRIX  $\bar{B}$**   
**FOR KELLY AFB, TX, CEILING PROBLEM**

$w_0 = 1$	$\bar{C}_0 =$	$\begin{bmatrix} 1.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.3125 & 0.0000 & 0.0000 & 0.0000 \\ 0.2292 & 0.0000 & 0.0000 & 0.0000 \\ 0.4583 & 0.0000 & 0.0000 & 0.0000 \end{bmatrix}$
$w_1 = \frac{0.8327^n}{1.8481}$	$\bar{C}_1 =$	$\begin{bmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.5292 & 0.4797 & -0.6007 \\ 0.0000 & 0.3518 & 0.3189 & -0.3993 \\ 0.0000 & -0.8810 & -0.7987 & 1.0000 \end{bmatrix}$
$w_2 = \frac{0.2885^n}{1.6715}$	$\bar{C}_2 =$	$\begin{bmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.6706 & -0.9562 & 0.0209 \\ 0.0000 & -0.7012 & 1.0000 & -0.0219 \\ 0.0000 & 0.0307 & -0.0438 & 0.0010 \end{bmatrix}$

As the data in Table 6 or the first row of the Crout auxiliary  $\bar{D}$  in the preceding section confirm,  $\bar{C}_0$  indeed turns out to be the matrix of climatology, i.e., the predictor means:\*

\*As explained above, only in cases like the Kelly example, where each Markov state corresponds to exactly one binary variable, can we expect to find exact agreement between the classical Markov and equivalent Markov models. In models containing more than one category of variable (e.g., visibility as well as ceiling), the equivalent Markov model neglects certain Boolean combinations of predictors that would form Markov states in the classical model. Depending on the additivity characteristics of the variables selected for inclusion in the model, this neglect of Boolean combinations has a greater or lesser effect on the faithfulness with which the equivalent Markov model reproduces the behavior of the classical mode. One immediate consequence of the lack of exact correspondence between the equivalent and classical techniques is seen in  $\bar{C}_i$ , which will not give the predictor means exactly except in simple cases like the Kelly example. In general, the climatological means will be well approximated by  $\bar{C}_0$ , however. In fact, any significant departure of  $\bar{C}_i$  from climatology is evidence of non-additivity and gives us warning to examine closely the model's performance on independent data.



$$\bar{z}_{0,t_0} = 1.0000 \quad \bar{z}_{1,t_0} = 0.3125 \quad \bar{z}_{2,t_0} = 0.2292 \quad \bar{z}_{3,t_0} = 0.4583$$

Although the matrix  $\bar{B}$  is 4 x 4 in size, it contains one dependent row (and column). Accordingly, the number of non-zero eigenvalues is 4-1 = 3, and there is no hierarchical form  $\bar{C}_3$ .

How long, i.e., to how great a value of  $n$ , should we retain terms  $w_1\bar{C}_1$  and  $w_2\bar{C}_2$ ? If we require accuracy to the third decimal place in  $\bar{B}^n$ , then any variation in the third decimal place of the weighting factors  $w_i$  is necessarily significant. In this case, such a requirement results in keeping the first-order term  $\bar{C}_1$  until  $n$  exceeds 34, and the second-order term  $\bar{C}_2$  until  $n$  exceeds five. Thus, in obtaining powers of  $\bar{B}$  beyond the first few, substantial computational savings are possible by abbreviating the matrix polynomial to be evaluated. More important, we see that the first-order mean departure term has in this case a noticeable effect for at least 34 hours, while the second-order mean departure term becomes negligible after five hours. In other words, if our criterion for convergence to climatology is agreement in the third decimal place, we can say the system converges to climatology in roughly 36 hours. This can be thought of as the "time to stationarity"  $\tau$ . Only during this time can the Markov model forecast probabilities other than the long term climatological expectation. Beyond  $\tau$ , the model forecasts only climatology.

There likely exist better criteria than agreement to an arbitrary number of decimal places for establishing  $\tau$  and for deciding when terms in the matrix polynomial become negligible. One such criterion would be to retain only such terms  $w_i\bar{C}_i$  for a particular forecast period  $n \Delta t$  that contribute significantly to improving the Brier P-score. Thus the cutoff for each term could be obtained by verifying the equivalent Markov model on independent data for forecasts of length  $n \Delta t$ . Trial forecasts could be made and verified using first-order terms only (0,1), then using first- and second-order (0, 1,2), then (0, 1, 2, 3), etc., until added terms no longer significantly improve the P-score. We note that there is no need to verify the zero-order term alone, since the P-score for a forecast of pure climatology can be calculated analytically (Brier and Allen, 1951):

$$P_{\text{climo}} = 1 - \sum_{j=1}^P \bar{z}_j^2$$

where  $\bar{z}_j$  is the mean of the  $j$ th dummy variable, and where there are  $P$  such dummies. Obviously it is not possible to calculate analytically the Brier P-scores for the classical or equivalent Markov models themselves, since these scores depend on the representativeness, robustness and inherent predictive skill of the particular model developed. Good models will produce good P-scores, and bad models will do badly. Nor even is it possible to prescribe analytically the relative improvement in P-score brought about by retention of each hierarchical form. If the improvement could be calculated, then we could obtain the actual P-scores by using  $P_{\text{climo}}$  as the baseline P-score. Since we have already shown logically that the absolute P-scores cannot be determined analytically, it follows that the relative improvement in  $P$  is likewise not obtainable by analytical means.

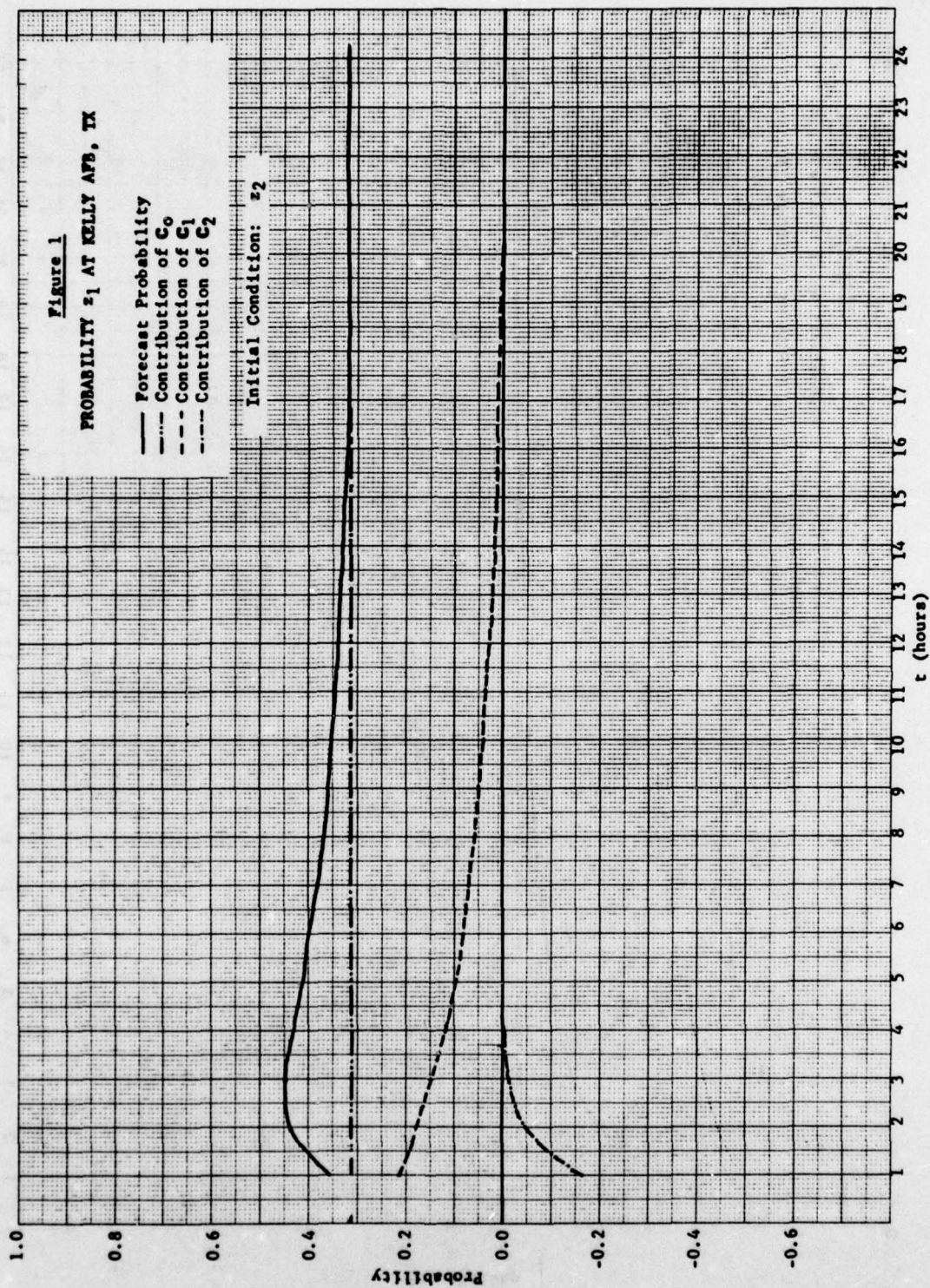
Whatever the exact method of obtaining the time to stationarity  $\tau$ , we can always regard  $\tau$  as an indirect measure of the potential skill of a Markov or equivalent Markov model. In general terms, the skill of a forecast technique is its ability to improve over some baseline objective forecast method such as random chance (forecasting the weather by dice), persistence (forecasting the weather will not change), or climatology (forecasting the mean). Using climatology as our no-skill baseline, we can see that a Markov prognostic scheme that converges to climatology in a time  $\tau < \Delta t$  will never make anything other than a climatological forecast. Such a scheme can never exhibit skill relative to climatology. In our Kelly example, on the other hand, the probability generating function  $\bar{B}^n$  does not converge to climatology to the third decimal place for about 36 hours. In other words, the model has 36 hours in which to make forecasts that differ somewhat from climatology, for better or for worse depending on the effectiveness of the model. The longer the time to stationarity  $\tau$ , the greater is the "skill opportunity" in the prognostic model. This is not to say, of course, that models with long  $\tau$  will necessarily be more skillful than those with short  $\tau$ . The Kelly equivalent Markov model is a case in point. Despite its relatively long  $\tau$ , this model would not likely show much skill in a test using independent data, since it was developed from an inadequate dependent data set. Nevertheless, if two Markov models are comparable in other respects, it can be expected that the one with longer time to stationarity  $\tau$  will exhibit greater skill than the other relative to climatology.

A hierarchical matrix analysis such as that shown in Table 8 can sometimes be used to gain added physical insight into the operation of the prognostic technique in terms of the atmospheric variables whose behavior is being modeled. Consider, for example, the third row of the hierarchical matrix  $\bar{C}_2$  in Table 8. This column contributes to the forecast probabilities only if the observed ceiling is  $z_2$ , i.e., 3,000 - 15,000 ft. In cases where  $z_2$  is observed, the effect of element (3,3) of  $\bar{C}_2$  is to give added weight to persistence of  $z_2$  in the short range, whereas elements (2,3) of  $\bar{C}_2$  and (4,3) of  $\bar{C}_1$  sharpen the probabilistic persistence forecast by decreasing the likelihood of weather change. Since  $\bar{C}_1$  is felt considerably more strongly than  $\bar{C}_2$  because of its larger weighting factor, the effect of (4,3) of  $\bar{C}_1$  dominates (2,3) of  $\bar{C}_2$ , and the second most likely short range event (after persistence, the most likely) is deterioration of the weather from  $z_2$  to  $z_1$ . Note that the heavy short range weight on persistence (3,3) of  $\bar{C}_2$  must yield after several hours to a favored deterioration of the weather, shown by the balance between (2,3) and (4,3) of  $\bar{C}_1$ . Only the greater *a priori* probability of good weather, shown in (4,1) of  $\bar{C}_1$ , prevents  $\bar{C}_1$  from making a "landslide Prediction" of  $z_1$ , the lowest ceiling. Nevertheless,  $\bar{C}_1$  is sufficiently strong to tilt the forecast probabilities in the direction of  $z_1$  for several hours (see the six hour forecast in Table 7 and the displayed forecast probabilities in Figure 4).

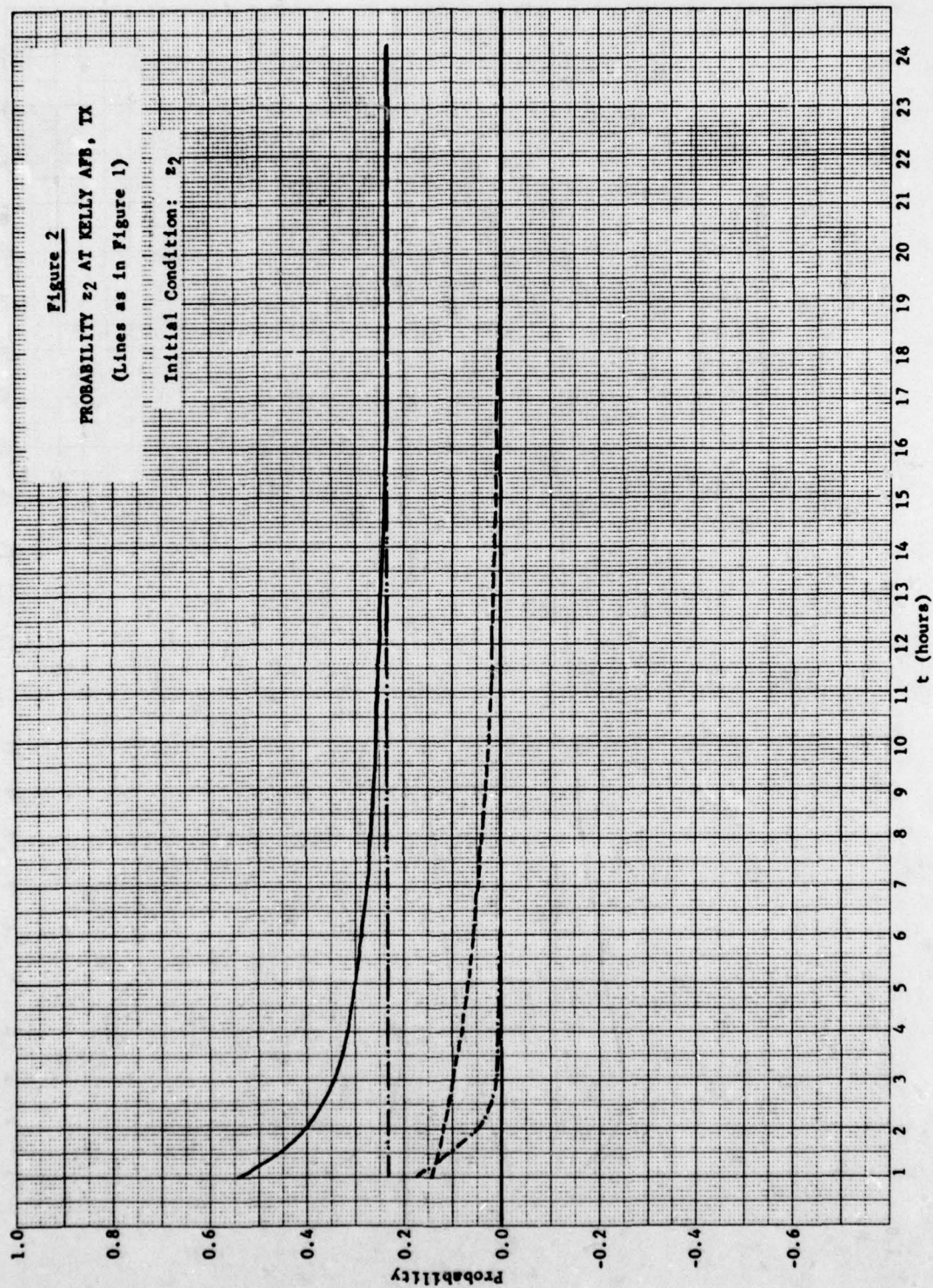
This sort of analysis is shown in graphical form in Figures 1 - 4, which display the forecast probabilities as a function of time for ceiling categories  $z_1$ ,  $z_2$  and  $z_3$  based on an initial condition of  $z_2$ . Moreover, in Figures 1 - 3, the probability contribution due to the hierarchical matrix forms  $\bar{C}_0$ ,  $\bar{C}_1$  and  $\bar{C}_2$  are also shown as a function of time. Clearly, the transient effect of persistence, which causes  $z_2$  to be the most probable category for the first 1 1/2 hours, is mostly due to  $\bar{C}_2$ , the most "short lived" hierarchical form. The forecast of lowering ceiling, which holds from  $t = 1.5$  to  $t = 8.5$ , is due principally to  $\bar{C}_1$ , with the decreasing probability of  $z_1$  in the first two hours being due to  $\bar{C}_2$ . The convergence of the forecast probabilities toward climatology  $\bar{C}_0$  operates at all times and is seen to dominate beyond 3 hours.

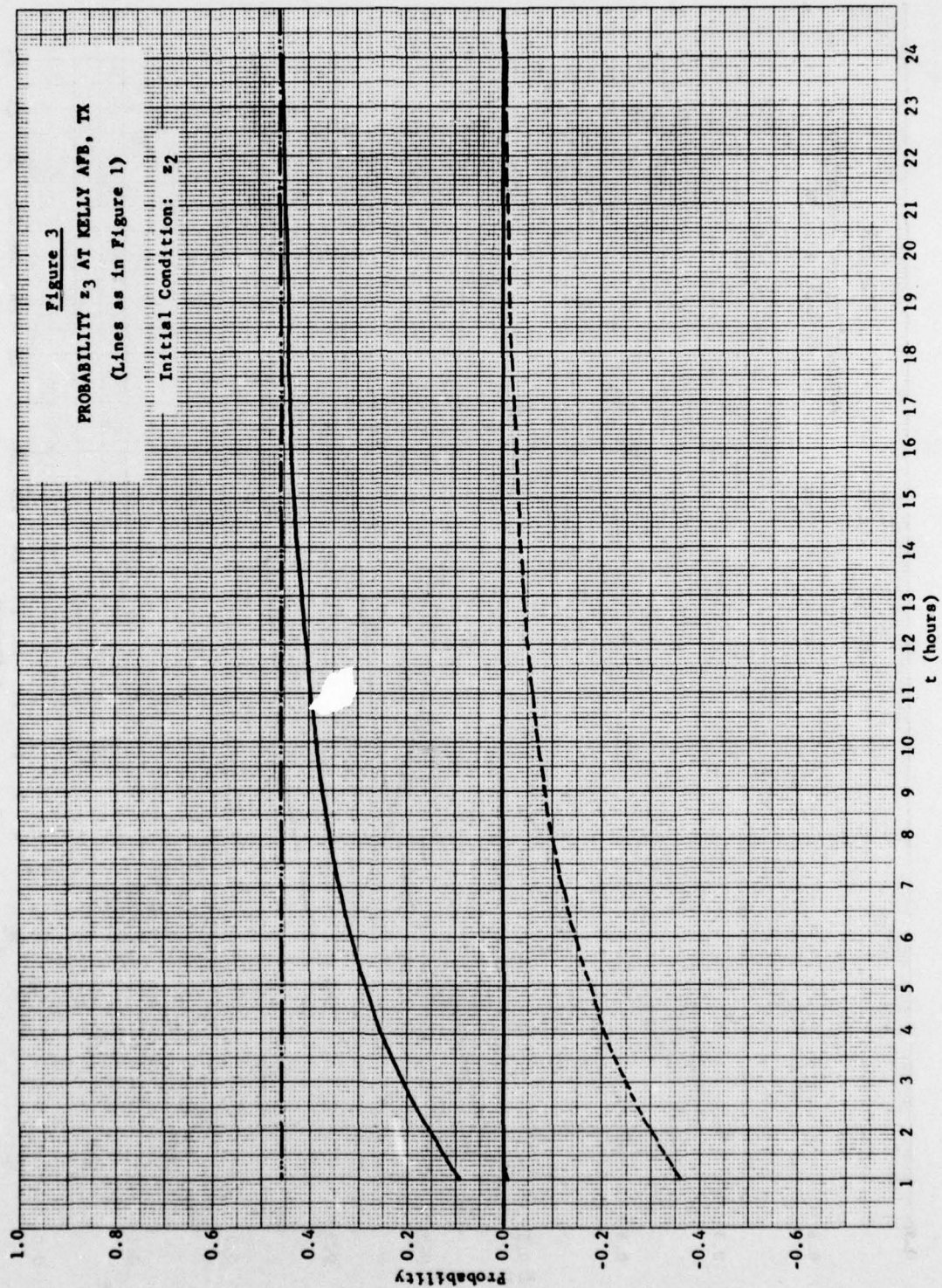
Thus we see in  $\bar{C}_2$  very short range "probability forces" at work favoring persistence. In the present case, these forces are operating on a time scale of several hours. In  $\bar{C}_1$  we see longer range "forces" at work over time scales of approximately a half day. It is from this source that the predicted deterioration of the weather emerges. Finally, we see in  $\bar{C}_0$  that climatology, operating most strongly beyond 12 hours, exerts an influence toward improving the ceiling.



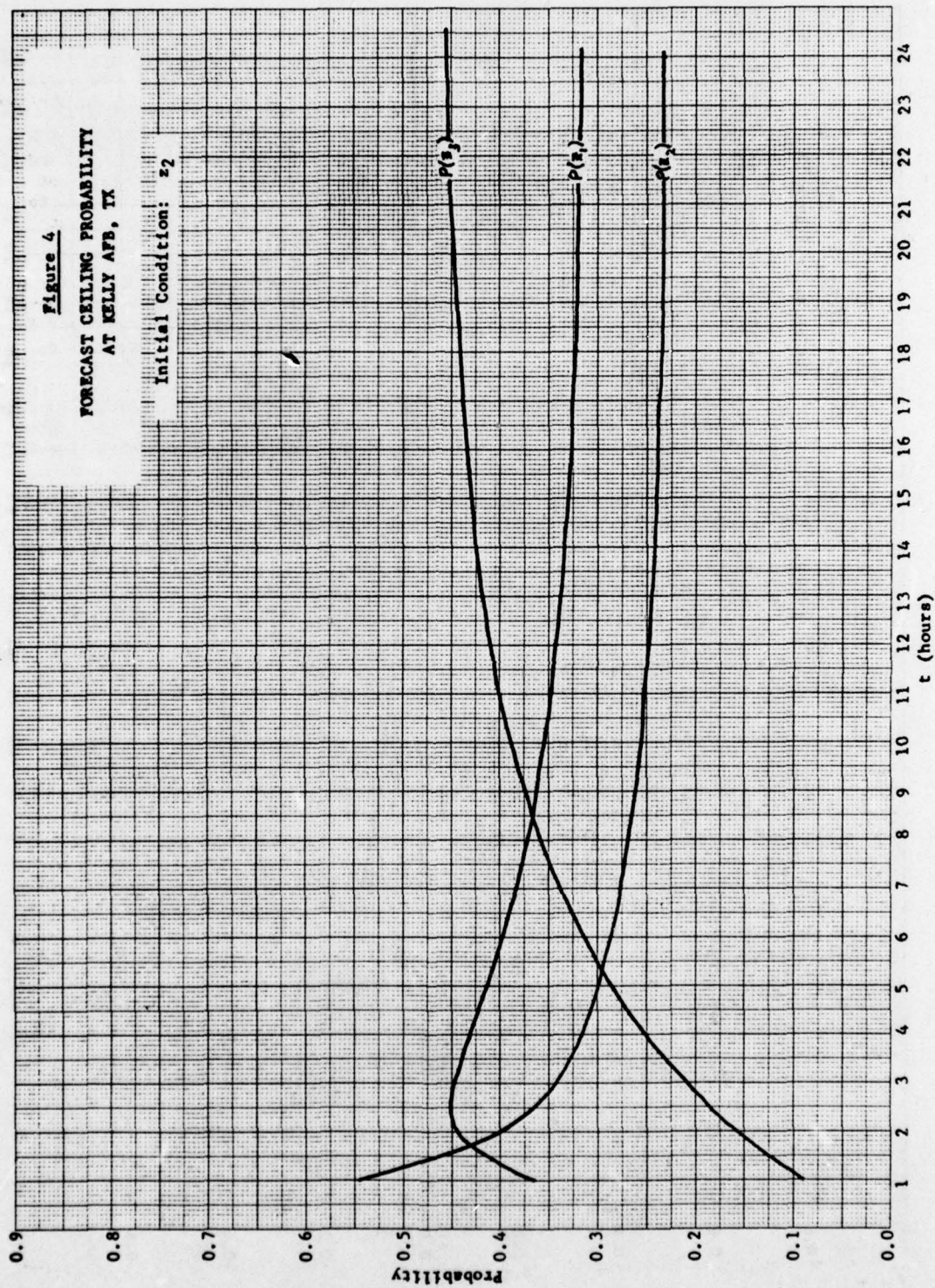














#### 14. The Ornstein-Uhlenbeck Process: Continuous Variates.

The classical Markov and equivalent Markov models require discretization of predictors and predictands into several classes, which we refer to as dummy or binary variables. Some atmospheric variables (such as the "present weather" elements, tornado, hail, thunder, etc.) are quite naturally treated in terms of dummy variables, which may be either "on" or "off." Other atmospheric elements such as temperature, pressure and the like are distributed continuously, and any effort to categorize them involves some loss of information. The classical Markov and equivalent Markov models trade decreased resolution in some of the variates for the important ability to include more than one variable in the prediction scheme. These additional predictors, if selected appropriately, impart increased prognostic skill to the classical and equivalent Markov models, particularly in forecasting the weather beyond 3 - 4 hours in the future.

For some purposes, particularly forecasting very short range weather changes on the order of minutes to 1 - 2 hours, the decreased resolution in the variate to be forecast might not always be remunerated by the increased forecasting skill provided by use of additional predictors. This is physically reasonable. We would naturally expect the probability density function of the visibility 15 minutes from now to be influenced more by the present visibility than by some other predictor such as dewpoint depression or even wind speed. Under these circumstances, the prognostic scheme used to make the 15-minute forecast might do better to make use of a thorough observation of the present visibility than to incorporate a host of additional predictors. In a case such as this, a Markov process involving a continuous variate might prove more skillful than either the classical or equivalent Markov techniques.

Gringorten (1966, 1968, 1971, 1972) has adapted for meteorological use a special class of the Markov chain called the Ornstein-Uhlenbeck process in which a single, continuous variate serves as both predictor and (through a time lag) predictand. Hering and Quick (1974) have employed this model with remarkable success in forecasting the atmospheric extinction coefficient  $\sigma$  ( $m^{-1}$ ) for 15, 30, 60 and 180 minutes in the Air Force Geophysics Laboratory's Mesonet Experiment. The success of this single station forecasting method warrants its discussion here.

Assuming a first-order Markov process, in which the outcome of trial  $n$  depends at most on the outcome of trial  $n-1$ ,\* in the stationary Ornstein-Uhlenbeck process the normalized variable  $y$  (having mean of zero and variance of 1) exhibits serial correlation given by

$$\rho_{n \Delta t} = e^{-an\Delta t}$$

where  $\rho_{n \Delta t}$  is the correlation between observations separated by an interval of time  $n \Delta t$  ( $\Delta t$  is the unit in which time is measured, e.g., 1 hour), and  $\rho$  is the so called serial correlation coefficient or autocorrelation coefficient between observations  $y(t)$  and  $y(t + n \Delta t)$  separated by time interval  $n \Delta t$ , which is any multiple of the basic unit of time  $\Delta t$  characteristic of the problem. The constant  $a$  need not be expressly determined, as it does not appear in the final equation.

Let  $\rho_0$  be the correlation coefficient between observations  $y(t)$  and  $y(t + \Delta t)$  separated by one unit of time  $\Delta t$ . Then,

$$\rho_0^{\Delta t} = e^{-a\Delta t}$$

If the process is Markov, then the correlation coefficient between observations  $y(t)$  and  $y(t + n \Delta t)$  separated by an arbitrary interval of time  $n \Delta t$ , which may be fractional, is given by

$$\rho = \rho_0^{n\Delta t}$$

Under the Ornstein-Uhlenbeck stochastic process, the value  $y_t$  of the continuous variate at time  $t$  is related to an earlier value  $y_0$  by the expression,

$$y_t = \rho y_0 + \sqrt{1 - \rho^2} p$$

where  $p$  is a normalized probability and where  $\rho$  is the correlation between  $y_t$  and  $y_0$  separated by the interval  $n \Delta t$ .

\*Hering (1977) reports that experimentation with higher order Markov processes involving the outcome of  $n-2$ ,  $n-3$ , etc., in the Ornstein-Uhlenbeck context did not produce significant improvement in the performance of the forecast scheme.

The Ornstein-Uhlenbeck model presented here has been tested and given extensive use in the Air Force Geophysics Laboratory's Mesonet Experiment, where it has been used both as guidance for subjective forecasters and as a control against which to measure the skill of various forecasting methods. As applied in the Mesonet Experiment, the model forecasts the extinction coefficient  $\sigma$  ( $m^{-1}$ ) in the form of the normalized variable  $y$ , where at initial time  $t_0$ ,

$$y_0 = k \ln \sigma_0 + f$$

The extinction coefficient at initial time  $t_0$  is  $\sigma_0$ , and  $k$  and  $f$  are coefficients that depend on time of day and season. Details are available in Tahnk (1975). The unit of time  $\Delta t$  in this application is 1 hour, so time is given as

$$t = n \Delta t$$

in hours (0.25 hours to 3 hours). Therefore, the Ornstein-Uhlenbeck model of a first-order Markov process is

$$y_t = y_0 e^{\rho t} + \sqrt{1 - (\rho e^{\rho t})^2} p$$

where  $p$  is a normalized probability that a certain threshold  $y_c$  will be exceeded and is obtained indirectly from the climatological record as explained in Gringorten (1972). The autocorrelation  $\rho$  is also determined from data but may be "tuned" to improve model performance.

The model equation, as written, produces a forecast of the most probable normalized extinction coefficient  $y_t$  for time  $t$  based on exponential decay of the autocorrelation coefficient with time. But because  $y$  is a normalized variable with mean of zero and variance of one, tables of the normal probability integral can be consulted or the normal distribution integrated numerically to obtain the probability of exceeding selected operationally significant thresholds. Once the "constants" have been determined, this model requires only an initial, uncategorized observation  $y_0$  of the parameter being forecast in order to estimate the most probable future value  $y_t$  and various "exceedance" probabilities. If  $y_0$  is observed continuously, the computer can make available continuous forecasts of  $y_t$ .

Hering and Quick (1974) report, based on the first year of the model's use in the Mesonet Experiment, that the model proved difficult for conventional forecasters to beat in 15-minute, 30-minute, 1-hour and 3-hour forecasts, especially in the 15- and 30-minute forecasting. Forecasters equipped with Mesonet data could make about a 10 percent improvement in Rank Probability Score over the Markov model in 15- and 30-minute forecasting but could not sustain this improvement at 1 and 3 hours. In 3-hour forecasting, the Markov model actually beat the conventional forecasters, who were denied Mesonet information. Percentage improvement in Rank Probability Score above the score produced by forecasts of pure climatology was 71, 60, 39 and 23 percent for 15-minute, 30-minute, 1-hour and 3-hour forecasts, respectively. The Markov probabilities were much less biased toward pessimism than were the subjective forecasts. Because the Markov model tends to dampen a perturbation with time as the model gradually "forgets" its initial state, the Markov probabilities were not as sharply cast as were the subjective probabilities.

A model such as this is limited by its lack of "additional" predictors, i.e., elements other than that being forecast. The simple exponential decorrelation that forms the basis of the model almost never produces a forecast of worsening weather. Under these circumstances, it is interesting to speculate how much improvement in forecast performance could be realized by introducing models containing more meteorology than the Ornstein-Uhlenbeck scheme provides. Tahnk (1975) reports the results of a test in which the performance of two regression-based models was compared with that of the Ornstein-Uhlenbeck model. One of the regression models was a REEP prediction scheme not much different from the REEP-based equivalent Markov model discussed earlier in this chapter. The other competing model was based on classical multiple linear regression equations with continuous variables, the predictors being selected by a stepwise scheme. Neither of the competing schemes was limited to single-station predictors, and in fact both regression schemes heavily chose network predictors\* in preference to single station predictors during model development.

In tests using independent data, the REEP-based technique proved much more skillful than the Ornstein-Uhlenbeck model in 15-, 30- and 60-minute forecasting, producing Heidke skill scores of 6.8%, 19.2% and 13.4% relative to Ornstein-Uhlenbeck for the respective forecast periods indicated. The stepwise regression scheme using continuous variables did about as well as the REEP technique at 15-minute forecasting but showed negative skill relative to the Ornstein-Uhlenbeck model at 30 and 60 minutes.

\*See section 16 of this chapter.



Stepwise did better than REEP at forecasting the onset of worsening conditions but did so at the expense of a high onset false alarm rate. Both regression models did better than the Ornstein-Uhlenbeck at forecasting onsets. Not unexpectedly, all three models found forecasting improving weather a much easier task than forecasting deterioration. Nevertheless, the stepwise regression scheme hardly did better than the Ornstein-Uhlenbeck model at this task, whereas the REEP technique picked up almost twice as many of the improving situations as did either of the other methods.

Tahnk (1975) made a special study of the usefulness of the REEP technique in cases of radiation fog only and found that in this more difficult situation REEP was less skillful than the Ornstein-Uhlenbeck model at 15-minute forecasting (-4.1% Heidke skill score). At 30 and 60 minutes, however, REEP beat Ornstein-Uhlenbeck by a wide margin (25.6 and 38.9%, respectively).

In a final experiment, Tahnk (1975) limited the REEP technique to selection of single station predictors only, i.e., Mesonet observations for the Hanscom AFB reservation itself. The rederived REEP equations were then tested relative to the Ornstein-Uhlenbeck model using independent data. The test was not restricted to radiation fog cases. The restricted REEP model showed large negative skill relative to the Ornstein-Uhlenbeck technique for all forecast periods. It thus appears important, at least in the mesoscale context, to allow the REEP model to select network predictors if these are available. At the synoptic scale, use of network predictors may not offer correspondingly large improvements in forecasting skill, although at least some gain is generally achieved by including network predictors. Chapter 10 of this volume presents a synoptic scale forecasting experiment in which the equivalent Markov or REEP technique is limited to single station predictors yet shows appreciable forecasting skill.

## 15. Ancillary Models.

### a. General.

More often than not, the user of weather information needs more than a simple forecast,

$$\Pr(z_0, z_1, z_2 \dots z_p)_t$$

of the probability of the weather at some future time  $t$ . Reconnaissance mission planners, tactical commanders and even construction project superintendents are likely to want to know, for example, "What's the probability the weather will be good 24 hours from now and will stay good for six hours after that?"\*

In fact, a well posed question such as this defines the weather element of the customer's mission success indicator (MSI), provided an objective meaning can be found for the customer's term, "good weather." More important, the manner in which the question is posed gives away the secret of its solution. Let us write the question graphically:

What is the probability that ...

$$\left\{ \begin{array}{l} \text{..the weather will be} \\ \text{good 24 hours from now..} \end{array} \right\} \text{ and } \left\{ \begin{array}{l} \text{..the weather will stay} \\ \text{good for 6 hours after} \\ \text{that..} \end{array} \right\} \quad ?$$

We can write this in a slightly more mathematical form, using  $\Pr(M)$  as the probability that the weather will be good enough for the mission to proceed.

$$\Pr(M) = q(W=G \text{ at } t_1 = t_0 + 24) \cdot p(W=G \text{ at } t_1 \text{ thru } t_7 | W=G \text{ at } t_1)$$

where both  $q$  and  $p$  are probabilities,  $W$  represents the "weather,"  $G$  stands for "good," and

$$t_n = t_1 + (n - 1) \Delta t$$

for  $\Delta t = 1$  hour.

We say that the customer's question gives away the solution to his problem because both the question (see the graphical form) and its mathematical statement are in two parts. We need both the forecast probability that the weather will be good at  $t_1$ , i.e.,

$$p(W=G \text{ at } t_1)$$

\*This is simply one example of the many such questions that can be posed under actual operational circumstances.



and the continuous run probability, i.e.,

$$p(W=G \text{ at } t_1 \text{ thru } t_7 | W=G \text{ at } t_1)$$

of good weather for six more hours.

The models described in this chapter and elsewhere in the volume address themselves for the most part to the classical forecasting problem of obtaining  $p(W=G \text{ at } t_1)$  for some future time  $t_1$ . The second part of problems such as that presented above is often best solved by methods different from but related to those used for the first part. We can conveniently consider here only a few of the methods developed in the literature. We will call these ancillary models because they are best suited to the "second part" of weather support problems such as that described above. In keeping with the subject matter of this chapter, only single station methods of the Markov type will be treated.

b. Continuous Run Probabilities by Markov Process: Hot and Cool Spells or Wet and Dry Periods.

When we are interested in particular sequences of Markov states, such as the probability of a cool day (C) following four consecutive hot days (H), i.e.,  $\Pr(H_1, H_2, H_3, C_4)$ , we need not construct the whole Markov transition matrix because we are interested only in certain elements of the matrix. Sakamoto (1970) found that under a first order Markov process, the probability of a particular sequence of Markov states  $a_i$  is given by

$$\Pr(a_1, a_2, a_3 \dots a_m) = q(a_1) p(a_2|a_1) p(a_3|a_2) \dots p(a_m|a_{m-1})$$

where the subscripts refer to the sequence number of the Markov trial and where  $q$  represents the initial probability. In the particular case of three consecutive hot days followed by a cold day, the algorithm is

$$\Pr(H_1, H_2, H_3, C_4) = q(H_1) p(H_2|H_1) p(H_3|H_2) p(C_4|H_3)$$

Obviously, the conditional probabilities  $p(a_j | a_i)$  are first order Markov transition probabilities, given in this case by

		<u>Final State</u>	
		<u>C</u>	<u>H</u>
<u>Initial State</u>	<u>C</u>	$p_{cc} = p(C C)$	$p_{ch} = p(H C)$
	<u>H</u>	$p_{hc} = p(C H)$	$p_{hh} = p(H H)$

for the two-state hot/cold system.

In a first order Markov model, the outcome of trial  $n$  is presumed to depend only on the outcome of trial  $n - 1$  immediately preceding. In a second order Markov model, trial  $n$  is affected not only by  $n - 1$  but also by  $n - 2$ . Sakamoto (1970) shows that a second order Markov chain for the same example would be written as

$$\Pr(H_1, H_2, H_3, C_4) = q(H_1) p(H_2|H_1) p(H_3|H_2, H_1) p(C_4|H_3, H_2)$$

where the conditional probabilities can involve as much as two days prior to the day of interest. A third order chain would be

$$\Pr(H_1, H_2, H_3, C_4) = q(H_1) p(H_2|H_1) p(H_3|H_2, H_1) p(C_4|H_3, H_2, H_1)$$

involving the conditional probability that day  $m$  will be cold given that days  $m - 1$ ,  $m - 2$  and  $m - 3$  were hot.

Sakamoto found that the manner in which the Markov states, "cold day" and "hot day," were defined affects the performance of the resulting Markov chain in modeling the continuous run probabilities (e.g.,  $\Pr(H_1, H_2, H_3, C_4)$ ). In fact, choice of a threshold used to define a hot or cold day proved to be a factor in determining the suitability of a particular order of the Markov chain model. Inappropriate thresholds may force the use of a second order chain, an eventuality that might be avoided by use of thresholds better suited to the prediction problem at hand.

Use of arbitrary temperatures (e.g., 86°F, 89°F, ...) to define hot and cold days proved unsuitable in the context of either a first order or a second order Markov chain. More effective were definitions based on a certain number of degrees above a moving weekly mean maximum (e.g., 3°F, 6°F, 9°F, ... above the moving mean). Even with relative thresholds such as this, it was shown that the value of the relative threshold (e.g., 6° or 9°) dramatically affects the performance of the Markov model and the suitability of a particular order of the chain.

By resorting to relative criteria for "hot" and "cold" days, Sakamoto found he was able to use a first order Markov chain to approximate the probability of hot and cold spells of various lengths. Use of second order Markov models produced a slight improvement in the resulting probabilities, but the small gain in precision was not sufficient to justify the added computational expense in developing and applying the more complex models.

Similar work in applying first order Markov chains to cool and hot spells has been reported by Caskey (1964) and Spiegel (1966), who indicated the first order model applied well enough. Caskey (1964) applied the model of Gabriel and Neumann (1962), discussed below, to the problem of calculating the probability of cold spells in London. The model equation is

$$\Pr(C, n) = [1 - p(C|C)] p(C|C)^{n-n_0} \quad n \geq n_0$$

where  $\Pr(C, n)$  represents the probability of  $n$  days of cold (C) weather, i.e., a cold spell of  $n$  days' duration. The conditional probabilities  $p(C|C)$  represent the likelihood that a day will be cold given that the previous day was cold, a constant independent of  $n$ . Note that the exponent is not the number of days of the cold spell  $n$  but rather the number of days by which the cold spell exceeds the baseline length  $n_0$ , given by Caskey as four. This is because the probability that a cold spell of length  $n_0$  will continue  $k$  further days is  $p(C|C)^k$ , where

$$k = n - n_0$$

Other authors have considered whether sequences of wet and dry days can adequately be represented by first order Markov chains. Caskey (1963), Feyerherm and Bark (1965) and Gabriel and Neumann (1962) report that the first order Markov process is generally adequate for this purpose, although in some cases it has not been completely satisfactory. Spiegel (1966) and Feyerherm and Bark (1965) tested higher order Markov models. While these models did show slight improvement over the first order Markov chain, the improvement did not seem to justify the greater computational expense involved in development and application of the more complex models.

#### c. Probability of Precipitation or No Precipitation in an Interval of Time.

If we consider that there are two kinds of days, dry (D) and wet (W), then the probability  $\Pr(W, n)$  that precipitation will occur at some time during an interval of  $n$  days can be given in terms of the probability  $q(W, n-1)$  of precipitation in  $n-1$  days and the conditional probability  $p(W, n | D, n-1)$  of a wet  $n$ th day following a period of  $n-1$  dry days.

As is the case in many combinatorial problems, this problem is best solved by considering the complementary probabilities, i.e., the probability of dry weather. The probability  $\Pr(D, n)$  of  $n$  consecutive dry days is simply one minus the probability  $\Pr(W, n)$  of precipitation occurrence on any day among the  $n$  days. Mathematically,

$$\Pr(D, n) = 1 - \Pr(W, n)$$

Likewise, the conditional probability of a dry day following a period of  $n-1$  dry days is simply one minus the conditional probability  $p(W, n | D, n-1)$  of a wet day following  $n-1$  dry days, i.e.,

$$p(D, n | D, n-1) = 1 - p(W, n | D, n-1)$$

and the marginal probability  $q(D, n-1)$  of  $n-1$  dry days is one minus the marginal probability  $q(W, n-1)$  of  $n-1$  wet days:

$$q(D, n-1) = 1 - q(W, n-1)$$

Thus the probability of  $n$  consecutive dry days is

$$\Pr(D, n) = q(D, n-1) p(D, n | D, n-1)$$



In terms of complementary probabilities, this is

$$1 - \Pr(W, n) = [1 - q(W, n-1)] \cdot [1 - p(W, n | D, n-1)]$$

Expanding and rearranging the equation yields the recursive form,

$$\Pr(W, n) = q(W, n-1) \cdot [1 - p(W, n | D, n-1)] + p(W, n | D, n-1)$$

If the process is first order Markov, the conditional probability  $p(W, n | D, n-1)$  that the  $n$ th day will be wet following  $n-1$  dry days is influenced solely by whether the day immediately preceding day  $n$  is dry or wet. For consistency with our earlier notation, let us say that for the first order Markov process,

$$p(W, n | D, n-1) = p(W | D)$$

Under these circumstances, the model equation becomes

$$\Pr(W, n) = q(W, n-1) \cdot [1 - p(W | D)] + p(W | D)$$

Caskey (1963) shows that this equation reduces to

$$\Pr(W, n) = 1 - [1 - q(W)] \cdot [1 - p(W | D)]^{n-1}$$

where  $q(W)$  is the marginal probability of any day being wet.

If this model equation is to be applied to real problems, then values of  $q(W)$  and  $p(W | D)$  must be obtained from the historical data. Two choices are open. Either we can estimate both  $q(W)$  and  $p(W | D)$  directly from the data; or we can obtain  $p(W | D)$  and  $p(W | W)$  from the data and then obtain  $q(W)$  from the identity of Gabriel and Neumann (1962):

$$q(W) = p(W | D) \cdot [1 - p(W | W) + p(W | D)]^{-1}$$

Consistent with our notation,  $p(W | W)$  is the conditional probability that a day will be wet given that the previous day was wet.

In applying this model to actual rainfall data, Caskey (1963) found that sequences of wet and dry days can be represented adequately by the first order Markov chain.

A first order Markov model suitable for estimating the probability of wet and dry spells of arbitrary length, the probability of exactly  $s$  wet days among  $n$  days following a wet or dry day, the probability of  $s$  wet days among any  $n$  days, and the probability of a weather cycle of  $n$  days has been formulated by Gabriel and Neumann (1957, 1962) and Gabriel (1959), who found the model fits Tel Aviv daily rainfall data quite well.

In developing the model, Gabriel and Neumann (1957) reasoned that the lengths of wet and dry spells conform to a so called "geometric" distribution (Feller, 1950, p 217) to a good degree of approximation. Under the geometric distribution, the probability that a positive, integral-valued random variable  $X$  equals a particular integer  $k$  is given by

$$\Pr(X=k) = [1 - p] p^{k-1}$$

where  $k = 1, 2, 3, \dots$  and  $p$  is a conditional probability. We can particularize this to the problem of wet and dry spells. We reason that  $p(W | W)$  is the probability of a wet day followed by another wet day, i.e., the probability of two wet days in sequence. Likewise, the probability of three wet days in sequence is

$$p(W | W) p(W | W)$$

and of four wet days is

$$p(W | W) p(W | W) p(W | W)$$

Generalizing, it is apparent that the probability of  $k$  wet days in a row is

$$p(W | W)^{k-1}$$

which can be interpreted as the probability that a wet day will be followed by  $k-1$  wet days. The probability  $p(D | W)$  that a wet day will be followed by a dry day is the complement of  $p(W | W)$ , so

$$p(D | W) = 1 - p(W | W)$$



We may think of a wet spell as a sequence of  $k$  wet days terminated by at least one dry day. Thus the probability of a wet spell of exactly  $k$  days is the probability of  $k$  wet days followed by a dry day, or

$$\Pr(W, k) = p(D|W) \cdot p(W|W)^{k-1}$$

or

$$\Pr(W, k) = [1 - p(W|W)] \cdot p(W|W)^{k-1}$$

which is identical with Feller's geometrical distribution, given above. The same reasoning can be used to obtain the probability of a dry spell of length  $k$ :

$$\Pr(D, k) = [1 - p(D|D)] \cdot p(D|D)^{k-1}$$

Obviously, we are working with a Markov transition matrix:

		<u>Current Day</u>	
		<u>Dry (D)</u>	<u>Wet (W)</u>
<u>Previous Day</u>	<u>Dry (D)</u>	$p_{DD} = p(D D)$	$p_{DW} = p(W D)$
	<u>Wet (W)</u>	$p_{WD} = p(D W)$	$p_{WW} = p(W W)$

so that

$$p(D|D) + p(W|D) = 1$$

$$p(D|W) + p(W|W) = 1$$

In such a model, the probability of rainfall  $i$  days after a wet day is

$$\Pr(W, W+i) = q(W) + [1 - q(W)] \cdot [p(W|W) - p(W|D)]^i$$

where  $q(W)$  is the marginal probability of a wet day, given by

$$q(W) = \frac{p(W|D)}{1 - p(W|W) + p(W|D)}$$

The probability of rainfall  $i$  days after a dry day is, by analogy,

$$\Pr(W, D+i) = q(W) + q(W) \cdot [p(W|W) - p(W|D)]^i$$

It is apparent that both  $\Pr(W, W+i)$  and  $\Pr(W, D+i)$ , the probabilities of rainfall  $i$  days after a wet or dry day, respectively, both converge to  $q(W)$ , the marginal probability of a wet day, with increasing  $i$ . This is the Markov process gradually "forgetting" its initial state. The marginal probability may be estimated from

$$q(W) = p(W|D) \cdot [1 - p(W|W) + p(W|D)]^{-1}$$

as shown above.

Defining a weather cycle as a combination of a wet spell and an adjacent dry spell, Gabriel and Neumann (1962) found the probability of a weather cycle of length  $n$  days is

$$\Pr(C, n) = p(W|D) \cdot [1 - p(W|W)] \cdot \left\{ \frac{[1 - p(W|D)]^{n-1} - p(W|W)^{n-1}}{1 - p(W|D) - p(W|W)} \right\}$$

The probability of exactly  $s$  wet days among  $n$  days following a wet day is

$$\Pr(sW|W+n) = p(W|W)^s [1 - p(W|D)]^{n-s}$$

$$\sum_{c=1}^{s+1} \binom{s}{c} \binom{n-s-1}{b-1} \left[ \frac{1 - p(W|W)}{1 - p(W|D)} \right]^b \left[ \frac{p(W|D)}{p(W|W)} \right]^c$$

where

$$c_1 = \begin{cases} n + \frac{1}{2} - |2s - n + \frac{1}{2}| & \text{for } s < n \\ 0 & \text{for } s = n \text{ (and sum then involves only this term)} \end{cases}$$

and where  $a$  and  $b$  are the least integers not smaller than  $(1/2)(c-1)$  and  $(1/2)c$ , respectively. Similarly, the probability of exactly  $s$  wet days among  $n$  days following a dry day is

$$\Pr(sW|D+n) = p(W|W)^s [1 - p(W|D)]^{n-s} \cdot \sum_{c=1}^{c_0} \frac{[s-1]!}{[b-1]!} \frac{[n-s]!}{[a]!} \left[ \frac{1 - p(W|W)}{1 - p(W|D)} \right]^a \left[ \frac{p(W|D)}{p(W|W)} \right]^b$$

where

$$c_0 = \begin{cases} n + \frac{1}{2} - |2s - n - \frac{1}{2}| & \text{for } s > 0 \\ 0 & \text{for } s = 0 \text{ (and sum then involves only this term)} \end{cases}$$

The probability of  $s$  wet among any  $n$  days is

$$\Pr(sW|n) = q(W) \Pr(sW|W+n) + [1 - q(W)] \Pr(sW|D+n)$$

Gabriel and Neumann (1962) suggest that the close fit of their simple, first order Markov model to daily rainfall in Tel Aviv implies the distribution of wet and dry spells is not "periodic" or "harmonic" to any appreciable extent in the location and at the scale considered. Furthermore, in order to fit the Markov model's requirement for the independence of day  $n$  rainfall from that on any day earlier than  $n-1$ , the daily rainfall -- at least at Tel Aviv -- must be among those weather elements that "decouple" or decorrelate rapidly. Presumably, certain other weather elements might not decouple as fast as the daily rainfall and might therefore not be modeled as well by a first order Markov process. Other weather elements might decouple faster, perhaps allowing use of a model in which the weather on day  $n$  is assumed to be entirely independent of weather on all previous days. The Bernoulli process (Bernoulli trials), whose outcomes conform to the binomial distribution, might then be a more appropriate model than the Markov chain.

#### 16. Development of the Regression Equations: Generalized vs. Localized Operators.

We have seen that an equivalent Markov prognostic model can be developed by using a multiple linear regression scheme to produce a probability generating function  $\bar{B}$  in matrix form. Screening regression would logically be used to select the predictors included in the multiple linear regression relationships whose coefficients define  $\bar{B}$ .

Left unresolved in this analysis has been the fundamental question of what shall be included in the data set to be fed to screening regression to establish the multiple linear regression relationships. Classically, the approach has been to use only predictand data from the individual station for which forecasts are desired. The resulting multiple linear regression prediction equations (and probability generating function  $\bar{B}$ ) then apply only at the station in question. There might in addition be an effort to stratify the data by season, thus obtaining several sets of statistical prediction equations, each peculiar to a particular station and season. Such sets of equations are said to be localized operators in that their applicability is local in space, time, or both. By far the greater number of local forecast studies have been made in this way (George, 1960), presumably because their authors felt better skill could be obtained by particularizing the predictands to one time and place.

This concept has been successfully challenged by Harris, Bryan and MacMonegle (1963), who advanced the concept of a generalized statistical operator as a viable alternative to the localized operator. A generalized operator is a statistically derived specification or prediction equation that has general applicability throughout different times of year and at different geographical locations (Harris, et al, 1963).



Under the generalized operator concept, data are not stratified by predictand station and season but rather are grouped by continent-size region or possibly into a single, worldwide sample. Under those circumstances, the prediction equations derived apply at any location in the region or world, not just at one point. It is well to point out that although the statistical prediction equations thus developed are generalized and apply over a large geographical region, the predictor values have to be computed separately for each point at which the equations are to be solved. Otherwise, of course, the scheme would predict homogeneous weather over the entire region.

The justification for the generalized operator concept is ultimately that the atmosphere obeys the same physical laws over Tennessee that it obeys over Texas, Tahiti or Tibet. There is no seasonal or geographical variability in the universal physical laws governing the motion, composition and state of the atmospheric fluid. Therefore, if statistical relationships can somehow be made to capture the essential physics of the atmosphere rather than be dominated by temporal and spatial variation in the absolute value of some of the atmosphere's variables, then the prognostic applicability of those relationships will not be limited in space and time.

Indeed, generalized operators are not new. Dynamical weather prediction models are among the best examples of generalized deterministic operators. In these models, known physical laws are expressed mathematically in the form of a somewhat simplified model of the atmosphere. Analogous to these generalized hydrodynamical equations in the deterministic domain are the generalized statistical operators in the domain of uncertainty. Limitations in numerical weather prediction models and error in meteorological observations now prevent dynamical, deterministic prediction of surface weather elements such as ceiling and visibility. Under these circumstances, statistical operators in a mixed dynamical-statistical prediction scheme can help bridge the gap between model capabilities and user expectations, producing probabilistic forecasts of surface weather elements based on dynamical predictions of upper air patterns.

From a practical point of view, there exist several reasons for preferring generalized statistical operators over their localized equivalents:

- The meteorologist is frequently required to forecast the weather at a location having either no historical weather data at all or having a period of record inadequate for preparation of classical, localized operator forecast studies. Part of the job of the military meteorologist is to prepare himself to forecast for Bare Bases, remote sites and forward areas that he never heard of before being asked to forecast for them.
- Where there is a sudden demand for forecasts for many locations--as there would be in wartime--the derivation of a host of localized statistical operators may require more time than is available or involve too great a cost.

For these reasons, it is well to proceed as far as possible with the generalized operator concept. To do so, means must be found by which to subtract climatic and seasonal variability from the atmospheric variables entering the prediction scheme as predictors and predictands. It is not unreasonable to expect some success in this search to remove geography and season. Our science shows us that whereas the absolute values of many atmospheric variables may be influenced by seasonal climatic regime and location, it is often the relative measures such as changes and gradients in parameters like pressure or moisture that dictate the production of weather. Accordingly, it ought to be possible to treat or somehow transform these atmospheric variables such that their seasonal and geographical variability is removed. Simple methods of doing this are to use sea level pressure instead of station pressure, to give the temperature in terms of its deviation from the normal, or to use wind anomaly in lieu of raw wind data. We shall see below that better means are available for doing this sort of thing. For now, it suffices to know that we will have removed as much as possible of the geographical and seasonal variability of the parameters to be used as predictors and predictands in a generalized statistical operator.

How does the generalized operator work? As shown in Figure 5, one starts by removing the geographical and seasonal variability from the parameters  $X_j$  to be used as predictors. Sometimes in addition one normalizes the variables if one's prediction scheme requires normally distributed input data. The treated predictors are shown as  $X_j$  in Figure 5. Next the treated predictors are fed to the generalized statistical operator, which might be an equivalent Markov regression scheme such as the one discussed above, or perhaps a discriminant function. The output of the generalized operator is a probabilistic forecast  $P(Y_j)$  for each  $Y_j$ . A reverse transformation is used to obtain  $P(Y_i)$  for  $Y_i$  in the society's conventional units of measurement. The statistical operator is general in that it is used for all locations and times, whereas the transformations and inverse transformations are local, being peculiar to a time and place.



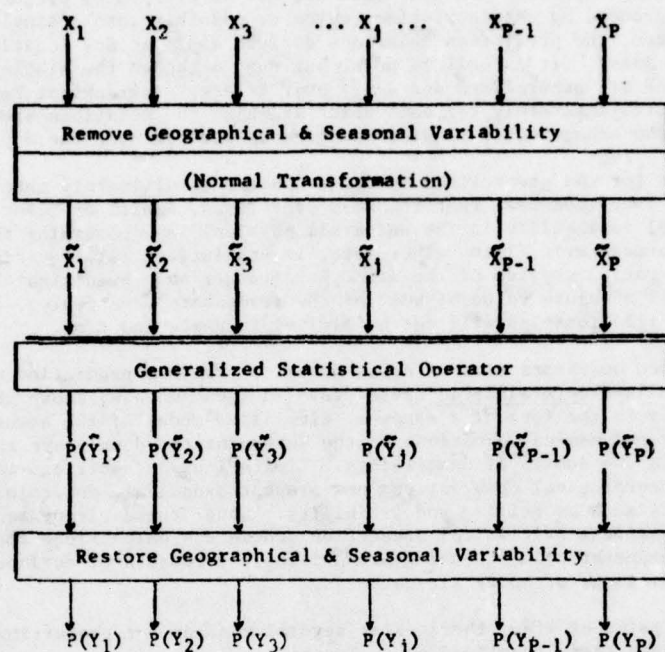


Figure 5. Use of a generalized statistical operator.

Generalized statistical operators such as this are of two types, the single station operator and the network operator. In short period prediction, persistence dominates and the observation at the station itself is found to provide a large share of the information necessary for successful prediction. Only the single station operator will be discussed to any great extent in this chapter.

Harris, Bryan and MacMonegle (1963) developed generalized statistical operators of both the single station and the network type for terminal forecasting at arbitrary locations over extensive geographical areas. The single station operators could use as predictors only data from the predictand location itself. Predictands included ceiling, visibility and total cloud amount in 2, 3, 6, 12, 18 and 24-hour forecasts. Predictors emphasized the primary Sutcliffe development equation terms, vorticity advection and thermal (thickness) advection. Developed from dependent data, the generalized operators were tested against a withheld, or independent data sample.

To remove seasonal and geographical variability from parameters used as predictors and predictands, Harris, et al, used transformed variables called standardized anomalies. These are essentially the anomaly  $A(X_{ij})$  of the  $i$ th observation of the  $j$ th variable  $X_j$ , where

$$A(X_{ij}) = (x_{ij} - \bar{x}_{.j})$$

$$\bar{x}_{.j} = \frac{1}{N} \sum_{i=1}^N x_{ij}$$

expressed in units of the variability of the variable. Specifically, the standardized anomaly  $S(X_{ij})$  of the  $i$ th observation of the  $j$ th variable is given by

$$S(X_{ij}) = \frac{x_{ij} - \bar{x}_{.j}}{\sigma(X_j)}$$

where  $\sigma(X_j)$  is the standard deviation of the  $j$ th variable. Both  $\bar{x}_{.j}$  and  $\sigma(X_j)$  are local to the time and place where  $X_j$  is observed.

Why divide by the standard deviation? This is done because the natural variability of the weather differs from one place to another. Probability theory and common sense tell us that "an anomaly is much more anomalous" if it represents a  $2\sigma$  deviation from the mean than if it constitutes a mere  $1\sigma$  deviation. Therefore, the standardized anomaly, which represents the mean departure in terms of the number of standard deviations by which it departs from the mean, is better related to the probability density of the variable  $X_j$  than the raw anomaly  $A(X_{1j})$  would be. Furthermore, the standardized anomaly is more free of seasonal and geographical variability than is the anomaly itself.

With normalized predictor and predictand data expressed in terms of standardized anomalies, Harris, et al, (1963) then used the method of screening regression to develop predictor/predictand relationships.

To test whether the generalized operator concept could compete effectively with the more traditional method of developing single-station relationships, Harris, et al, developed two sets of predictor/predictand relations, each constructed from a different dependent data set:

- The generalized operator, single-station relations were developed by combining normalized standard anomaly data from several nearby stations, with the data from one other station being withheld. For example, one data group included Syracuse, NY; Brattleboro, VT; Hanscom Field, MA, and Providence, RI, with Westover AFB, MA, being withheld.
- The localized operator, single-station relations were developed using only data from the station withheld from the generalized operator data set, i.e., Westover AFB.

The predictors and predictands selected from the generalized operator forecasting relations were also used in the localized operator equations.

With two sets of single station prediction equations developed--one generalized operator set and one localized operator set--both sets were used to make forecasts on independent data for the withheld station (i.e., Westover AFB). A simple persistence forecast was also run as a baseline for evaluation of forecasting skill. Two scores were used to compare performance of the competing forecast schemes: percentage of hits and the Heidke skill score. Prefigurance and postagreement were also computed.

In ceiling forecasts, both techniques beat persistence beyond about 6 - 8 hours, and the generalized single station operator performed about as well as the localized single station equation. The same picture held for visibility, except the statistical forecast techniques beat persistence at all forecast periods beyond 2 hours. In forecasting the total cloud amount, both single station techniques lost to persistence for all forecast periods, although the generalized operator method again did about as well (or as poorly) as the localized operator technique.

Harris, et al, felt that the power of their prediction scheme was not seriously limited by its use of a multiple linear regression equation to generate the forecast probabilities. Although the regression equation itself is assuredly linear, the predictands need not be. In fact, Harris, et al, (1963, 1965) used as one of their predictors the highly non-linear vorticity advection:

$$-\vec{v}_H \cdot \nabla_H \zeta = -u \frac{\partial^2 v}{\partial x^2} + u \frac{\partial}{\partial x} \left( \frac{\partial u}{\partial y} \right) - v \frac{\partial}{\partial y} \left( \frac{\partial v}{\partial x} \right) + v \frac{\partial^2 u}{\partial y^2}$$

Although prognostic techniques based on a network approach are strictly beyond the scope of a chapter on single station forecasting, it is well to point out that Harris, et al, (1963) tested generalized network operators as well as generalized single station operators. The network operators, except for forecasts less than about 3 hours, improved upon persistence in ceiling, visibility and total cloud amount forecasting. The network techniques were in general less skillful than single station methods for short range forecasts (less than about 3 - 6 hours), while at the longer forecast periods, the networks' ability to diagnose the advection of meteorological fields gave them the edge over single station techniques. Interestingly, although the single station approach had significantly fewer "hits" than persistence at all forecast periods in predicting total cloud amount, the network techniques had more hits than persistence for all such cloud amount forecasts except the 2-hour one. It appears cloud amount forecasting benefits more from having available a spatial field of information than does either ceiling or visibility forecasting.

In 1965, Harris, Bryan and MacMonegle extended their 1963 work using generalized operators. They retained their basic predictive scheme in which the probability of surface weather elements



was forecast by generalized statistical operators using as predictors only parameters derivable from 500, 700 and 1000 mb analyses. The earlier equations (Harris, *et al*, 1963) emphasized the primary Sutcliffe terms, vorticity advection and thermal (thickness) advection. The later work (Harris, *et al*, 1965) included as predictors such permanent effects as orography and coastal influence, the former providing a means of including forced vertical motion and the latter providing for moisture sources. Also included as predictors were time-of-day parameterizations of the incident solar radiation. In contrast to today's model output statistics (MOS), Harris, *et al*, employed a "perfect prog" concept, where predictors were taken from analyses rather than prognoses. In the 1965 work, five predictands were used: ceiling, visibility, total cloud amount, integrated operating condition (ceiling/visibility categories), and occurrence of precipitation. The generalized operators for each predictand were developed based on a sample of 16 stations spread evenly throughout the United States east of 100°W. Another sample of 15 stations for the same area was withheld from development and used as independent data in later tests of the method. Altogether, data from 31 stations were used either as a development sample or an independent test sample.

Three sets of prediction equations were prepared. The first set was permitted to use only the purely meteorological parameters used in the 1963 work, such as vorticity advection and thickness advection. The second set of equations was permitted to use these plus the orographic and coastal effects terms. In deriving the final set of equations, the screening regression scheme was also permitted to select radiation (time-of-day) parameters. Having three sets of prediction equations allowed the authors to examine the incremental prognostic value provided by the additive terms. It was found that the orographic and coastal effects predictors added significant information to the estimation of all predictands in one or more seasons of the year. Forecast results along the West Coast and along the eastern slopes of the Rockies were particularly improved by the orographic and coastal effects terms. The time-of-day or radiation parameters added even more significant information when permitted to enter the schemes for forecasting all elements except precipitation. In fact, time-of-day was selected more frequently than any other predictor in equations forecasting the visibility.

The prediction schemes of Harris, *et al*, (1965) improved upon "pure climatology" forecasts for all predictands and all seasons, with the greatest improvement occurring during fall and winter. The relations derived held up when applied to independent data. The prognostic schemes proved weakest in mountainous regions. The method using generalized operators was able to relate changes in upper air patterns used as predictors to changes in surface weather (predictands). The predictand fields could be analyzed much as weather maps themselves are. Features showed continuity in space and time and retained a good correspondence with upper air patterns. The statistical methods were able to produce realistically tight gradients separating large regions of fair skies from small areas of adverse weather.

In applying the method of generalized operators, there are three problems to be overcome.

First, predictors must be devised that adequately characterize the physical processes producing the weather elements being forecast. Parameters such as thickness advection, vorticity advection and pressure change are good candidates. The parameters devised as predictors must be feasible for computation on an operational basis.

Second, climatological biasing of the values of the predictor and predictand parameters selected must somehow be removed or the resulting prediction equations will be unduly specific to one place and time.

Third, physical characteristics of a particular locality may generate strong predictor/predictand relationships that are not generally valid over the region as a whole. These local effects must not be allowed to contaminate the generalized operator. It is generally possible to devise and add to the list of candidate predictors various parameterizations of local effects (e.g., station elevation, onshore component of the wind). Then, the generalized statistical relationship is developed from a sample of data including many stations. Under these circumstances, there should only be minimal contamination of the generalized operator by station-peculiar relationships.

#### 17. Single Station Forecasting Models in an Operational Setting.

Single station forecasting techniques such as those presented in this chapter are particularly feasible for operational application because of their simplicity. Once such a model is developed, the user can obtain a probabilistic forecast of all future states of the weather at his location simply by entering the conditions he observes at the time the forecast is made. No observations from outlying stations are needed. If the forecasting model is developed as a generalized operator, such that it has wide applicability over a large region, then the model constitutes a powerful, mobile, stand-alone weather support capability, accompanying the forecaster on a tactical operation, jumping with him into an assault zone, or moving with him from place to place as the command element



he supports is relocated. Provided forecasts are desired only for the forecaster's immediate vicinity, the models presented in this chapter have no need of weather communications. Nor do these models depend on uncertain assistance from centralized weather support facilities that may themselves be under attack or may be otherwise occupied with higher priority tasks.

A tactical scenario for single station Markov models is not difficult to imagine. Under extremely primitive circumstances, the forecaster could be equipped simply with a small book containing the several hierarchical matrix forms  $C_i$  needed to compute  $P^n$  in the equivalent Markov model. The actual computations could be done by hand, requiring at most 5-10 minutes to compute the probability of each weather element whose forecast is desired. Under more permissive circumstances, the forecasts could be prepared automatically by a microprocessor or a programmable hand calculator. The equivalent Markov model for forecasting the ceiling at Kelly AFB, for example, fits easily into today's rugged Hewlett-Packard 67 programmable calculator. In developed weather support facilities such as the Tactical Weather System (TWS) or Automated Weather Dissemination System (AWDS), today's small, highly reliable minicomputers would not be challenged at all by the simple computational tasks involved in making single station forecasts by the methods we have described.

Indeed, models such as this can form the basis of automated meteorological watch and short range forecasting to be performed at the highly automated weather stations of the future. At centralized weather support facilities such as the Air Force Global Weather Central (AFGWC), these models can be expected to bridge the gap between initial time and the period 12-24 hours later when numerical weather prediction models finally stabilize and begin to provide useful forecast guidance through model output statistics (MOS). With both AFGWC and the weather stations using the same prognostic models for at least the first 12 hours or so, compatibility of forecasts among the various facilities is insured.

#### 18. Summary and Conclusions.

We have seen that simple, yet powerful prediction methods of the Markov type can successfully be applied to the problem of forecasting the weather at a station, given an initial observation of the weather at that station only. This is single station forecasting by means of statistical prognostic models.

An equivalent Markov model due to Miller (1968) has been treated in detail and found to be comparable to the classical Markov model but much easier to develop and apply to practical forecasting problems. The same model is applied to a large scale prediction problem in chapter 10 of this volume. It is shown that the equivalent Markov model can make use of predictors produced by the model output statistics (MOS) method or on the other hand can be applied in the tactical context, where key data are often denied and single station methods are needed. In the context of weather support at developed facilities such as the minicomputer-equipped weather stations of the future, the equivalent Markov model is proposed as a means of generating probabilistic forecasts that bridge the gap between the current observation and the first useful numerical prognoses. Use of models such as this both at weather centrals and in the field would insure compatibility of forecasts.

Decision assistance to specialized users such as mission planners and command and control is shown to benefit from the use of so called ancillary models, which make use of the probabilistic predictions generated by the equivalent Markov technique.

It is shown that by using appropriate generalized operators it becomes possible to devise a very small number of single station forecasting methods that, taken together, are capable of generating single station (or network) forecasts for any location in the world, including those for which local climatology is not yet available. Thus a small repertoire of equivalent Markov schemes and an appropriate supporting system of ancillary models can adequately prepare a weather service to meet a wide range of peacetime and wartime weather support requirements, some of which may be difficult to anticipate and tedious to prepare for by other means.

## CHAPTER 8

### Nonlinear Prediction

by

John W. Louer

#### CONTENTS

1. Introduction.....	
2. Nonlinearity.....	
3. Screening Lattice Algorithm (SLAM) for Nonlinear Regression Estimation of Event Probabilities.....	
4. Applications of SLAM.....	
5. Conclusion.....	
6. References and Bibliography.....	

#### ILLUSTRATIONS

Figure 1. Conditional Distribution of $\bar{T}_1$ Given $\bar{T}_0$ .....	
Figure 2. Piecewise Estimation of Conditional Distribution.....	
Figure 3. Venn Diagram.....	
Figure 4. Cap, Cup, and Complementation Operators.....	
Figure 5. The Lattice of Divisors of 12.....	
Figure 6. The Boolean Lattice of Two Variables.....	
Figure 7. The Boolean Lattice of Two Variables in Zero-One Notation.....	

#### TABLES

Table 1. Relationship of Predictors $X_1$ and $X_2$ to Prob ( $Y=1$ ).....	
Table 2. Boolean Functions for Two Predictors.....	
Table 3. Logical and Operator.....	
Table 4. Logical or Operator.....	
Table 5. Exclusive or Operator.....	
Table 6. Boolean Predictors for Ceiling 200 feet ( $X_1$ ) and Visibility 1/2 mile ( $X_2$ ).....	

### 1. Introduction

In meteorology the ability to predict the weather at some time later utilizing the weather parameters that are measured at the present time is a goal that has been long sought by meteorologists. One problem is that there is uncertainty involved as to the relationships between the present and future weather parameters. Therefore, the use of probabilities facilitates the expression of these uncertainties. The various statistical methods of estimating prediction probabilities all have their weaknesses. One weakness that is common to several methods is the inability to account for non-linear relationships of a joint relationship type. This chapter addresses that problem.

The definition of the prediction problem in terms of the present weather parameters (predictors) and the weather parameters that we wish to predict at some time in the future (predictands) is as follows:

Predictors:  $\underline{X} = X_1, X_2, \dots, X_p, \dots, X_p$  (1)

Predictands:  $\underline{Y} = Y_1, Y_2, \dots, Y_g, \dots, Y_g$  (2)

where  $\underline{X}$  and  $\underline{Y}$  are vector quantities and the  $X_1, X_2$ , etc. represent present temperature, clouds, winds, etc. and  $Y_1, Y_2$ , etc. represent temperature, clouds, winds, etc. at a later time.

Using statistical methods, the objective is to predict  $f(\underline{Y}|\underline{X})$ , the conditional distribution of  $\underline{Y}$  given  $\underline{X}$ . For example, using linear regression, if one wanted to predict tomorrow's mean temperature  $\bar{T}_1$ , given today's mean temperature  $\bar{T}_0$ ,  $f(\bar{T}_1, \bar{T}_0)$  would be depicted as in Figure 1.

If the conditional distribution is known the solution is somewhat easier; if the distribution is not known then one can perform a piecewise estimation of  $f(\underline{Y}|\underline{X})$  as depicted in Figure 2.

Some of the statistical techniques that are available to predict  $f(\underline{Y}|\underline{X})$  are as follows (Miller, 1969):

a. Contingency methods: These methods are weak because cell frequencies become small or nonexistent when the number of predictors increases or the sample is small.

b. Stepwise regression (Miller, 1962) works well when the predictors and predictands are continuous and  $f(\underline{Y}|\underline{X})$  is Gaussian; this method is parametric.

c. Nonparametric Discriminant Analysis (Miller, 1962): Miller (1969) mentions that this method is very powerful, but computationally burdensome and therefore infeasible. Nonparametric refers to the fact that the distribution is unknown. (Siegel, 1956). Additivity is assumed to hold.

d. Regression Estimation of Event Probabilities (REEP), (Miller, 1964) is a stepwise regression procedure where the predictors and predictands are zero-one variables. It is powerful like Nonparametric Discriminant Analysis, but the calculations are not as burdensome and therefore it is not infeasible. It is also nonparametric. Additivity is assumed in REEP.

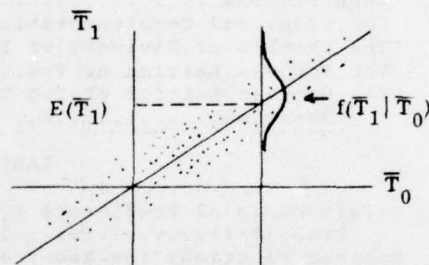


Figure 1. Conditional Distribution of  $\bar{T}_1$  (tomorrow's mean Temperature) Given  $\bar{T}_0$  (today's mean Temperature). or

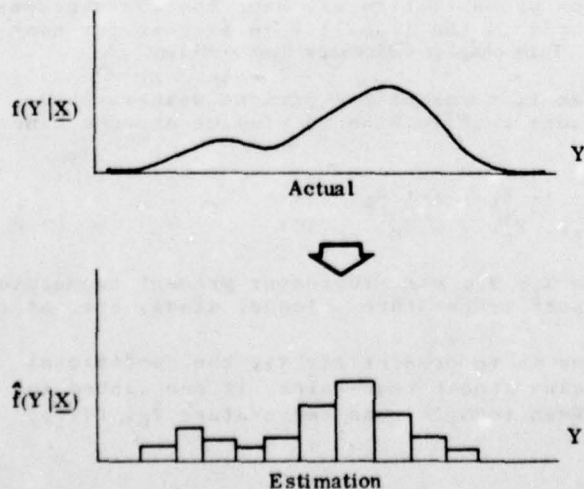
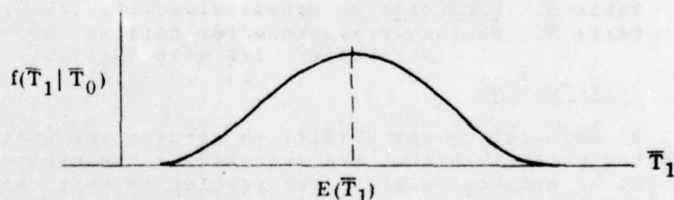


Figure 2. Piecewise Estimation of Conditional Distribution (Miller, 1969).



Two binary operations and an operation of complementation can be defined using the Venn diagram. The "cap" operation,  $\cap$ , equates to the intersection of the two subclasses or predictors and is also the Boolean logical and operation, symbolized by  $\cdot$ . The "cup" operation,  $\cup$ , or the union, equates to everything within the two sub-classes and is the Boolean logical or operation which is symbolized by  $+$ . The complementation operation is symbolized by a "—" and means everything within the population outside of the subclass. Figure 4 depicts with a Venn diagram these operations utilizing predictors  $X_1$  and  $X_2$ .

A set of  $N$  predictors has  $2^N$  possible Boolean functions. For the two predictors  $X_1$  and  $X_2$ , therefore, 16 possible Boolean functions exist. Table 2 depicts the 16 possible Boolean functions, their arithmetic equivalent, and their corresponding Venn diagrams (Miller, 1969).

The logical or and logical and Boolean operators can be defined as in Tables 3 and 4 respectively. The 13th function in Table 2 is usually defined as the exclusive or operator (Table 5).

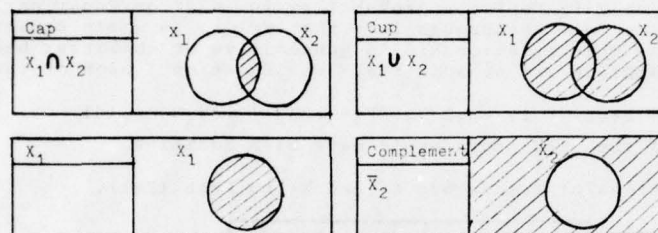


Figure 4. Cap, cup, and complementation operations (Flegg, 1964)

When  $N$  is large, any tests of the significance of each Boolean function with regard to a particular predictand would be prohibitive in time and effort. However, out of 16 possible Boolean predictors, only seven depict situations that add information. To illustrate this fact, assume that two predictors,  $X_1$  and  $X_2$ , are defined as follows:

$X_1$  - cloud ceiling <200 feet  
 $X_2$  - Visibility <1/2 mile

If we were concerned with predicting the probability that an airfield would be below landing minimums at some later time, these two conditions would very possibly be predictors.

Using the Boolean functions from Table 2, the physical explanations would be as given in Table 6. As can be seen from the table, only three of the Boolean predictors would not be redundant and would add any information beyond that provided by the REEP formulation. They are  $X_1$ ,  $X_2$ , and  $X_1 \oplus X_2$ . The remaining thirteen are derivable as linear functions of these three.

Miller (1969) used the fact that there is a partial ordering amongst the  $2^N$  possible Boolean functions to develop an algorithmic method of uncovering the joint predictive information that one is looking for and further reduce the testing for significance of the Boolean functions. This partial ordering is the lattice property.

Figure 5 depicts a non-Boolean lattice of the devisors of 12 and depicts the idea of partial ordering. All numbers connected by lines moving upward in the figure are devisors of the higher numbers. In a Boolean lattice the divisor property is not applicable, but a subset property can be depicted in a similar manner as in Figure 6. Each node on this lattice of the predictors  $X_1$  and  $X_2$  represents a subset of the nodes connected to it upwards in the lattice. For example, referring to the numbers of the Boolean functions in Table 2 which are shown on Figure 6, Function 11 is a subset of Functions 2, 13, and 5. These functions are then subsets of Functions 3 and 8, 3 and 9, and 8 and 9 respectively. Figure 7 depicts the same Boolean lattice in zero-one notation.

The basic principle underlying the SLAM is that if a Boolean node in the lattice is significantly related to a particular predictand, one must decide whether to proceed up or down in the lattice in order to obtain more information or more refined information. If the Boolean function is significantly related to the predictand, it may be reflecting more important sources of information based upon its position in the lattice (Miller, 1969).

## 2. Nonlinearity

One weakness is common to all of the statistical methods discussed earlier except the contingency methods; nonlinear information contained in the joint relationships among the predictors must be known *a priori*. In linear relationships the effects of the predictors are additive, while in nonlinear relationships the effects are not additive (Siegel, 1956). Tukey (1949) discusses nonadditivity and suggests a test for it.

An example of nonadditivity best illustrates this property. Suppose that one has the following equation:

$$\text{Prob } (Y=1) = .2 + .2X_1 + .2X_2 \quad (3)$$

where  $X_1$  and  $X_2$  are zero-one variables. Their relationship to the probability that  $Y=1$  is shown in Table 1. The effects of the two predictors in this table may have been determined by using contingency methods. If neither  $X_1$  nor  $X_2$  are "on" (equal one), i.e., neither occurred and are zero, the probability that  $Y=1$  is .2. Likewise, if either  $X_1$  or  $X_2$  is "on" the probability is .4 (from Equation 3). However, if both  $X_1$  and  $X_2$  are "on" the probability that  $Y=1$  is .9 while Equation 3 would give a probability of .6. This relationship is nonadditive or nonlinear because the effects of the predictors are nonadditive. Had Equation 3 been of the form

$$\text{Prob } (Y=1) = .2 + .2X_1 + .2X_2 + .3X_1X_2 \quad (4)$$

then the effects of the predictors would have been additive.

Table 1. Relationships of Predictors  $X_1$  and  $X_2$  to Prob ( $Y=1$ ).

		Predictor $X_2$	
		0	1
Predictor $X_1$	0	.2	.4
	1	.4	.9

## 3. Screening Lattice Algorithm (SLAM)

Miller (1969) developed an algorithm to overcome this weakness in REEP. This algorithm, the screening lattice algorithm (SLAM), can be used to find joint predictive information among the predictors.

The SLAM utilizes Boolean algebra and lattices to find the joint relationships. In order to use Boolean algebra, all of the predictors must be considered as zero-one variables. Since the REEP predictors are zero-one variables, this algorithm can easily be utilized in conjunction with REEP.

Flegg (1964) offers an explanation of the Boolean algebra and Boolean functions that the SLAM employs. Utilizing Boolean algebra one can study the significance of relations between classes (Flegg, 1964).

If one has two predictors,  $X_1$  and  $X_2$ , which are elements of the total population of predictors, they and their relationships can be depicted using a Venn diagram and they can be considered as sub-classes of the total population. This pair of predictors is shown in Figure 3.

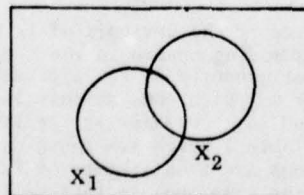







Figure 3. Venn Diagram

Table 2. Boolean Functions for Two Predictors (Flegg, 1964 and Miller, 1969)

Function Number	Binary Operation	Boolean Function	Arithmetic Equivalent	Venn Diagram
1	$x_1$	$x_1$	$x_1$	
2	$x_2$	$x_2$	$x_2$	
3	$x_1 \cup x_2$	$x_1 \oplus x_2$	$x_1 + x_2 - x_1 x_2$	
4	$x_1 \cap x_2$	$x_1 \odot x_2$	$x_1 x_2$	
5	$\bar{x}_1$	$\bar{x}_1$	$1 - x_1$	






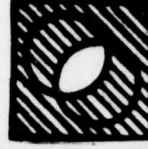


Function Number	Binary Operation	Table 2 (cont.) Boolean Function	Arithmetic Equivalent	Venn Diagram
6	$\bar{x}_2$	$\bar{x}_2$	$1 - x_2$	
7	$x_1 \bar{u} x_2$	$x_1 \oplus \bar{x}_2$	$1 - x_2 + x_1 x_2$	
8	$\bar{x}_1 u x_2$	$\bar{x}_1 \oplus x_2$	$1 - x_1 + x_1 x_2$	
9	$\bar{x}_1 \bar{u} x_2$	$\bar{x}_1 \oplus \bar{x}_2$	$1 - x_1 x_2$	
10	$x_1 \cap \bar{x}_2$	$x_1 \odot \bar{x}_2$	$x_1 - x_1 x_2$	
11	$\bar{x}_1 \cap x_2$	$\bar{x}_1 \odot x_2$	$x_2 - x_1 x_2$	

Table 2 (Cont.)






Function Number	Binary Operation	Boolean Function	Arithmetic Equivalent	Venn Diagram
12	$\bar{x}_1 \cap \bar{x}_2$	$\bar{x}_1 \odot \bar{x}_2$	$1 - x_1 - x_2 + x_1 x_2$	
13	$(\bar{x}_1 \cup \bar{x}_2) \cap (x_1 \cup x_2)$	$(\bar{x}_1 \oplus \bar{x}_2) \odot (x_1 \oplus x_2)$	$x_1 + x_2 - x_1 x_2 - x_1 x_2$	
14	$(\bar{x}_1 \cup x_2) \cap (x_1 \cup \bar{x}_2)$	$(\bar{x}_1 \oplus x_2) \odot (x_1 \oplus \bar{x}_2)$	$1 - x_1 - x_2 + x_1 x_2 + x_1 x_2$	
15	$\emptyset$	$\emptyset$	0	
16	I	$\Omega$	1	

Table 3. Logical or Operator (Miller, 1969).

$\Theta$	$X_2=0$	$X_2=1$
$X_1=0$	0	1
$X_1=1$	1	1

Table 4. Logical and Operator (Miller, 1969).

$\Theta$	$X_2=0$	$X_2=1$
$X_1=0$	0	0
$X_1=1$	0	1

Table 5. Exclusive or Operator (Miller, 1969).

Exclusive or	$X_2=0$	$X_2=1$
$X_1=0$	0	1
$X_1=1$	1	0

Table 6. Boolean Predictors for Ceiling <200 feet ( $X_1$ ) and Visibility <1/2 mile ( $X_2$ ).

Boolean Function	Physical Explanation	Remarks
1.	Ceiling <200 ft	
2.	Visibility <1/2 mi	
3.	Ceiling <200 ft <u>or</u> Visibility <1/2 mi	
4.	Ceiling <200 ft <u>and</u> Visibility <1/2 mi	
5.	Ceiling >200 ft	Redundant with No. 1. (Use 1 or 5)
6.	Visibility >1/2 mi	Redundant with No. 2. (Use 2 or 6)
7.	Ceiling <200 ft <u>or</u> Visibility >1/2 mi	
8.	Ceiling >200 ft <u>or</u> Visibility <1/2 mi	
9.	Ceiling >200 ft <u>or</u> Visibility >1/2 mi	Redundant with No. 4. (Use 4 or 9)
10.	Ceiling <200 ft <u>and</u> Visibility >1/2 mi	Redundant with No. 8. (Use 8 or 10)
11.	Ceiling >200 ft <u>and</u> Visibility <1/2 mi	Redundant with No. 7. (Use 7 or 11)
12.	Ceiling >200 ft <u>and</u> Visibility >1/2 mi	Redundant with No. 3. (Use 3 or 12)
13.	(Ceiling > 200 ft <u>or</u> Visibility >1/2 mi) <u>and</u> (Ceiling <200 ft <u>or</u> Visibility <1/2 mi)	
14.	(Ceiling >200 ft <u>or</u> Visibility <1/2 mi) <u>and</u> (Ceiling <200 ft <u>or</u> Visibility >1/2 mi)	Redundant with No. 13 (Use 13 or 14)
15.		Gives no meaningful information.
16.		Gives no meaningful information.



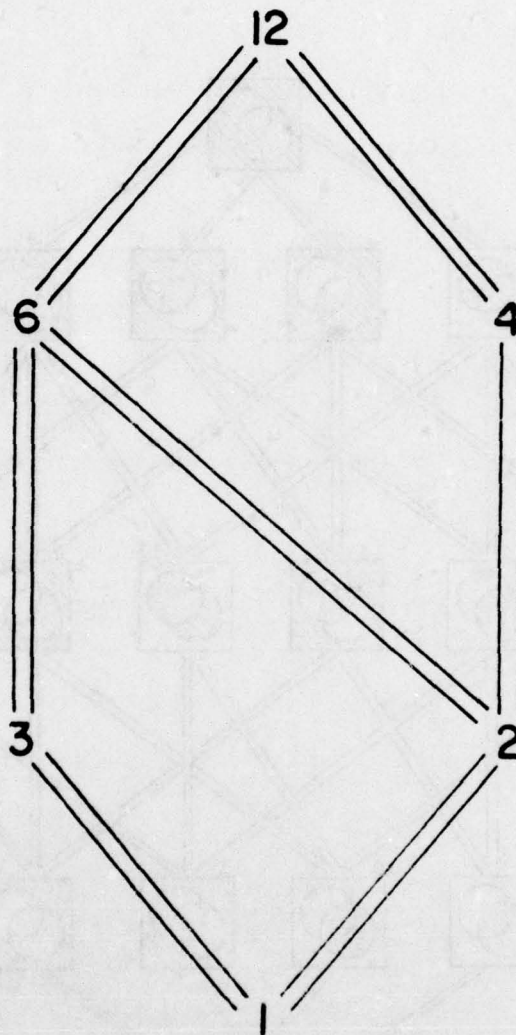


Figure 5. The Lattice of the Divisors of 12 (Miller, 1969).

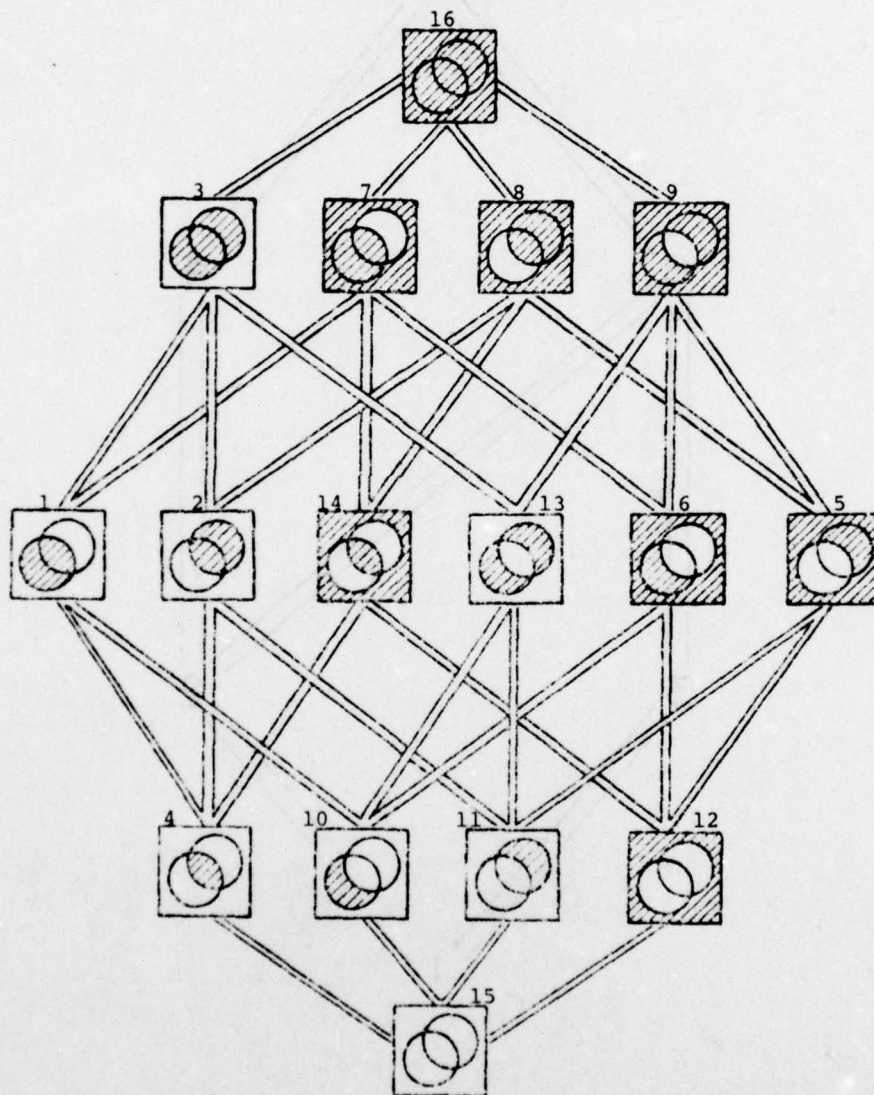


Figure 6. The Boolean Lattice of Two Variables (Miller, 1969)

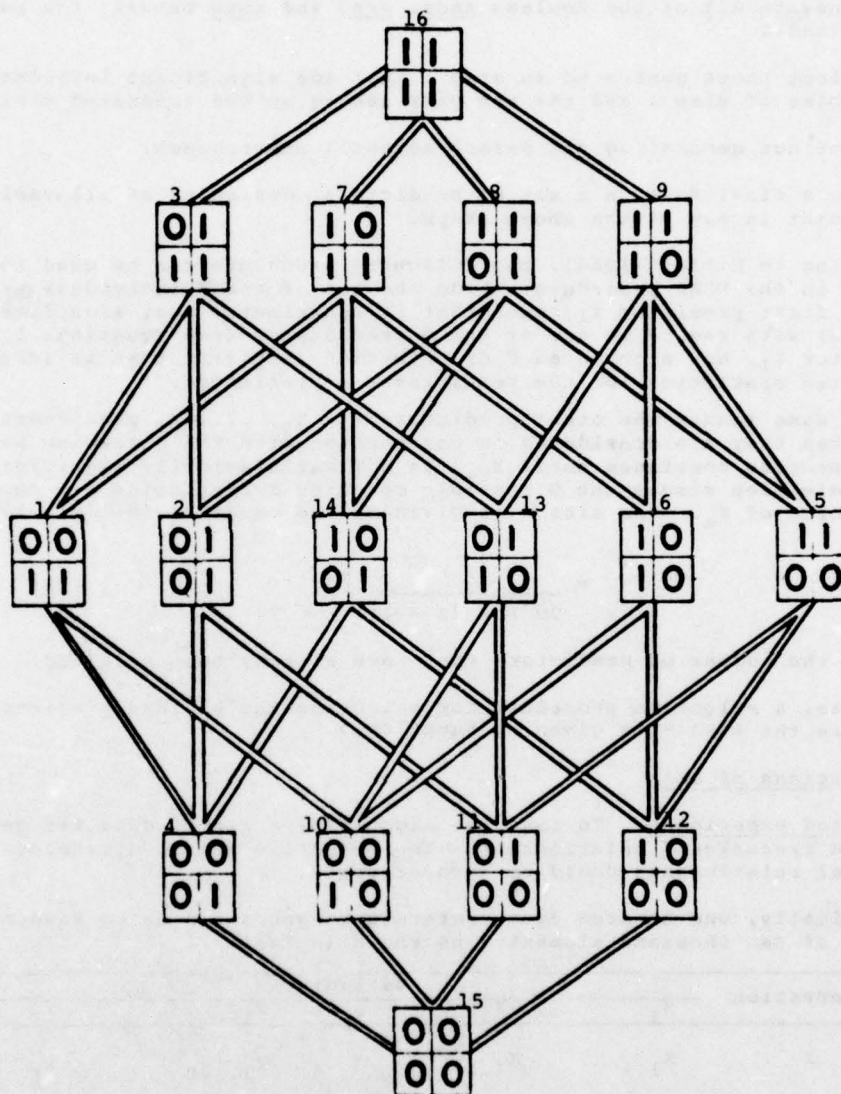


Figure 7. The Boolean Lattice of Two Variables in Zero-One Notation.



The resulting algorithm (SLAM) is as follows (Miller, 1969):

- a. Run the straight REEP on the data ( $X$  on  $Y_g$ ).
- b. Select the  $X$ 's which are significantly related (singly) to  $Y_g$ .
- c. Select significant exclusive ors among the unselected pairs of  $X$ 's from step b.
- d. Generate all of the Boolean ands, ors, and nots between the pairs selected in steps b and c.
- e. Select those generated in step d that add significant information over the REEP variables of step a and the raw pair making up the generated variable.
- f. Continue generating and selecting until convergence.
- g. Run a final REEP on a set of predictors consisting of all variables selected as significant in any of the above steps.

According to Miller (1964), the following procedure can be used to select the predictors in the REEP procedure: From the set of the  $P$  individual predictors, select the first predictor  $X_1$ , such that it contributes most significantly (better than chance) with regard to one of the  $G$  predictands (see Equations 1 and 2); the predictor  $X_1$ , has a computed  $F$  distribution statistic that is larger than any one of  $G$  computed statistics for the remaining  $P-1$  predictors.

In the same manner the other predictors  $X_2, X_3, \dots, X_r$  predictors can be selected when they are considered in conjunction with the preceding predictors. The selection process continues until  $X_{r+1}$  is not satisfactorily significant. At each of the  $r$  selection stages the  $G$  possible computed  $F$  statistics are compared with a critical value of  $F_\alpha$ . The size  $\alpha$  is given by the equation (Miller, 1964).

$$\alpha = \frac{1}{20 P - (S - 1)} \quad (5)$$

where  $S$  is the number of predictors that have already been selected.

Likewise, a selection procedure for selecting the Boolean predictors can use the  $F$  test where the size  $\alpha$  is given by Equation 5.

#### 4. Applications of SLAM

Simulated experiment: To test the algorithm, a set of data was generated which contained a preassigned relationship. The objective was to determine how closely the original relationship could be reconstructed.

Specifically, one hundred data vectors were generated using random numbers each consisting of ten thousand elements, as shown in Table 1.

Observation	Variables			
	$X_1$	$X_2$	$\dots X_{100}$	$Y$
1	$X_{1,1}$	$X_{1,2}$	$\dots X_{1,100}$	$Y_1$
2	$X_{2,1}$	$X_{2,2}$	$\dots X_{2,100}$	$Y_2$
3	$X_{3,1}$	$X_{3,2}$	$\dots X_{3,100}$	$Y_3$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
10000	$X_{10000,1}$	$X_{10000,2}$	$\dots X_{10000,100}$	$Y_{10000}$

Table 1

The individual observations are either zero or one. A one was assigned to an observation if a selected random number  $r$  ( $0 < r < 1.00$ ) was exceeded by the percentage desired for that variable. The 100 percentages were extracted from a random number table. Once the data matrix of Table 1 was constructed the following four Boolean variables were derived:

$$\begin{aligned} Z_1 &= (X_1 \otimes X_2) \otimes \bar{X}_3 \\ Z_2 &= (X_1 \otimes X_4) \\ Z_3 &= (X_2 \otimes X_4) \otimes (X_1 \otimes \bar{X}_5) \\ Z_4 &= X_2 \end{aligned}$$

Further, a linear function was constructed which has all the features of a REEP equation, i.e., estimates the probability that the predictand  $Y$ 's observation value is a one. For this simulated experiment the "REEP" equation is

Logical Form:

$$\text{Prob}(Y=1) = .399 + .239Z_1 + .137Z_2 + .257Z_3 - .384Z_4$$

or

Arithmetical Form:

$$\begin{aligned} \text{Prob}(Y=1) = & \begin{matrix} A_0 & A_1 & A_2 & A_3 & A_4 \\ .399 & - .144X_2 & - .239X_1X_2 & - .239X_1X_3 & + .137X_1X_4 \\ A_5 & A_6 & A_7 & A_8 \\ - .257X_1X_5 & - .239X_2X_3 & + .257X_2X_4 & + .239X_1X_2X_3 \\ A_9 & A_{10} \\ - .257X_1X_2X_4 & + .257X_1X_2X_4X_5 \end{matrix} \end{aligned}$$

From this equation each of the 10,000 observations of the  $Y$  vector were determined. Namely, a random number between zero and one was drawn and if its value was exceeded by the equation estimate a one was inserted in the  $Y$  vector for that observation otherwise a zero was inserted. The data were then subjected to a computer program which carried out the steps in the algorithm.

The results are shown below in arithmetical form.

		Constructed	Reconstructed
$A_0$	(Additive Constant)	.399	.399
$A_1$	$(X_2)$	-.144	-.149
$A_2$	$(X_1X_2)$	-.239	-.245
$A_3$	$(X_1X_3)$	-.239	-.245
$A_4$	$(X_1X_4)$	.137	.123
$A_5$	$(X_1X_5)$	-.257	-.248
$A_6$	$(X_2X_3)$	-.239	-.245
$A_7$	$(X_2X_4)$	.257	.277
$A_8$	$(X_1X_2X_3)$	.239	.245
$A_9$	$(X_1X_2X_4)$	-.257	.000
$A_{10}$	$(X_1X_2X_4X_5)$	.257	.000

It should be pointed out that the last two terms are not as different as might be interpreted at first glance. The simultaneous event  $X_1X_2X_4$  occurred 27 times in 10,000 while the event  $X_1X_2X_4X_5$  occurred 15 times. Thus, the lack of fit affects only 12 observations out of 10,000 since the coefficients on the last two terms nullify each other. This lack of fit may be attributed to the fact that statistical

significance was required of all Boolean functions selected. For the sample size used this small contribution failed to exceed the chosen significance level ( $\alpha = .05$ ).\*

It was concluded that reconstruction was performed to a satisfactory degree.

Further experiments: Three data bases were used to test SLAM. Each of these consisted of real world observations. The first of these, the Travelers automobile insurance data and the problem requiring solution, motivated the development of SLAM under the Travelers Environment and Man contract.

(1) Data Base: The Travelers Insurance Companies real-time data files.

Given: The predictors were the items recorded on the applications of 8,000 insureds, such as, sex, age, state of residence, use class, past driving record, occupation, etc. The predictand selected for study was the event that one or more claims would be submitted by the insured over the following twelve month period.

Objective: Predict the probability that the insured would submit one or more claims over the next twelve month period.

Solution: The equation determined from SLAM and the associated selected predictors were:

Travelers Insurance Companies Data

<u>Coefficients</u>	<u>Selected Predictors</u>
.073	(Additive Constant)
.145	NE Resident and Use Class 87
.137	Vermont Resident
.064	Age 68-72 and Small Land Area State of Residence
.068	NE Resident and Use Class 37
.056	Use Class 97
.044	Washington, D.C. Resident
- .044	Use Class 88
- .033	Idaho Resident
- .033	Low Population to Land Area State of Residence
.024	Occupation (Skilled, Unskilled, Technical, Factory Worker)
.024	State of Residence has Mandatory Inspection
.023	Use Class 87
.015	Accident Prior to 1966
- .015	Low Population Growth State of Residence

Independent Data Results:

	<u>Base</u>	<u>REEP</u>	<u>SLAM</u>
Information**	.2871	.2838	.2792
P score**	.1497	.1496	.1495
Hits	7348	7348	7348

Discussion: The predictors selected appear to be reasonable. The importance of each predictor can be interpreted as follows: When a predictor's condition is satisfied (its value is equal to unity) the amount of the corresponding coefficient is added to the additive constant (the first coefficient shown). Should the condition not be satisfied nothing is added. Thus, a New England resident in Use Class 87 will have .145 added to the probability that he will be submitting one or more claims within the twelve month period in question.

\*This 5% level of significance has been adjusted for the number of predictors screened. For a discussion of this point see Miller (1962).

\*\*Smaller values are preferred.



If an average claim is, say \$500, then a New England applicant satisfying Use Class 87 (unmarried; male; 21-24 years old, pleasure, work, or business use but not farmer; mileage im-material) should have his pure premium increased by \$72.50 to offset the added risk he represents to the insurance company. An independent test was made to determine the reliability of the selected predictors and of their coefficient. For an independent sample of 8,000 observations, the equation verified its reliability and skill over straight REEP.

- (2) Data Base: The Connecticut State Highway Department accident data.
- Given: The predictors were the characteristics recorded on 5,000 car-car accidents in Connecticut. These included such items as: location, hour of day, day of year, age of driver considered the victim, model year of automobiles, contributing factor, type of vehicle and violation.
- Objective: Predict the probability that the driver considered to be at fault was twenty years old or less.
- Solution: The equation was determined from SLAM and the associated selected predictors were:

Connecticut State Highway Department Data

<u>Coefficients</u>	<u>Selected Predictors</u>
.261	(Additive Constant)
-.218	Driver Under the Influence
.159	Hardtop and Darkness with Highway Illuminated
.142	Convertible
-.142	Convertible and Darkness with Highway Illuminated
.082	Vehicle Type Unknown (Other than Sports Car)
-.078	Model Year 1962
-.073	Daylight
.073	Daylight and 8 PM - 9 PM

Independent Data Results:

	<u>Base</u>	<u>REEP</u>	<u>SLAM</u>
Information*	.5421	.5241	.5214
P Score*	.3448	.3441	.3434
Hits	3189	3189	3189

- Discussion: Interpretations can be made regarding the importance of the selected predictors. For example, since most problem drivers are alcoholics and since few young drivers are alcoholics, it seems reasonable that .218 would be subtracted from the probability that the driver at fault is <20 given that the driver was under the influence. The independent sample of 5,000 cases confirms the stability and effectiveness of the SLAM procedure.
- (3) Data Base: United States Weather Bureau records for Hartford, Connecticut.
- Given: The predictors were the observed weather elements at forecast time, such as: temperature, cloud types, ceiling height, visibility, wind, humidity, weather (haze, fog, rain, snow, etc.), time of day and day of year. The predictor chosen was the event that there would be no weather one hour after forecast time.
- Objective: Predict the probability that there would be no weather (no haze, no fog, no rain, no snow, etc.) one hour in advance.

\*Smaller values are preferred.

Solution: The equation determined from SLAM and the associated selected predictors were:

Hartford, Connecticut Weather Data

<u>Coefficient</u>	<u>Selected Predictors</u>
.078	(Additive Constant)
.166	NO WEATHER
.045	CIG UNLIMITED
.058	CIG 1000-5000 feet
.284	VIS 15 miles
.266	VIS 10-14 miles
.191	VIS 7-9 miles
.161	JANUARY
.052	AUGUST
-.083	RLH 90-100%
-.064	WDR SE
.058	WSD 10-18 knots
.152	(NO WEATHER) (CTL SC)
-.161	(NO WEATHER) (CTL FS)
.005	(NO WEATHER) (CTU NS, ST)
-.161	(NO WEATHER) (JANUARY)
-.161	(CTL FS) (JANUARY)
-.266	(CTL FS) (VIS 10-14 miles)
-.116	(CIG 1000-5000 feet) (WSD 10-18 knots)
-.266	(VIS 10-14 miles) (LIGHT RAIN)
.005	(NO WEATHER) (WSD 2-8 knots) (CTU NS, ST)
.161	(NO WEATHER) (WSD 2-8 knots) (CTL FS)
.161	(NO WEATHER) (CTL FS) (JANUARY)
.266	(WSD 2-8 miles) (CTL FS) (VIS 10-14 miles)
.161	(WSD 2-8 miles) (CTL FS) (JANUARY)
.266	(CTL FS) (VIS 10-14 miles) (LIGHT RAIN)
.297	(CIG 1000-5000 feet) (DBT 71-80°) (VIS 3-6 miles)
-.161	(WSD 2-8 miles) (NO WEATHER) (JANUARY) (CTL FS)
-.266	(WSD 2-8 miles) (CTL FS) (VIS 10-14 miles)
	(LIGHT RAIN)

Discussion: A note of caution needs to be raised in this example. The selected predictors shown and their corresponding coefficients resulted from a very liberal test of significance. It turns out that no Boolean predictors showed significance at the usual 5% level over and above the predictors selected using straight REEP. The predictors shown are being presented only for illustrative purposes. The fact that none of the Booleans showed significance at the 5% level for such a short period weather forecast seems highly plausible. A more reasonable data base for uncovering Boolean predictors would have been for forecasts of longer range but unfortunately none was readily available.

5. Conclusion:

This report has presented one method for dealing with nonlinearity or nonadditivity in predicting meteorological events with statistical methods. This method, the SLAM, has the following features (Miller, 1969):

- a. The SLAM is nonparametric; no information about the underlying distribution is required.
- b. The SLAM is multivariate.
- c. The SLAM is nonlinear.
- d. The SLAM handles qualitative variables easily.

- e. The results are interpretable.
- f. Operational applications are easy to perform.
- g. The SLAM handles missing, erroneous, or incomplete data systematically.
- h. The SLAM processes large numbers of variables efficiently.
- i. The SLAM processes large numbers of observations efficiently.

Further tests of the SLAM in meteorological applications should be performed to determine its superiority or non-superiority over other methods.



## CHAPTER 9

### DELPHI TECHNIQUE

by Lt Col James W. Taylor

#### 1. Introduction.

For the systems analyst, such techniques as regression analysis, linear programming, and others too numerous to mention, are the tools of the trade. However, there is a class of problems for which the purely quantitative methods available to the analyst are either not appropriate, or have not yet been developed. For example, in the area of politics and strategy for military planning, the analyst must assess the probable intentions of the enemy by weighing and evaluating a tremendous amount of nonquantifiable data. An example is the development of the U.S. military role in space (Frye, 1968, p. 311). One method that might be used to address problems is the Delphi technique.

#### 2. What Is Delphi?

Developed by Olaf Helmer (1963), the Delphi technique is a methodology used to arrive at a consensus of opinion. By using the opinions of "experts" and by providing feedback, these experts are permitted to evaluate their own opinions in light of the opinions of others, and to make adjustments in their evaluations. After several iterations, a consensus of opinion is generally achieved, which, when quantifiable, has been found to be very accurate.

The originators of the Delphi technique believed it to be a possible method of forecasting future events. In fact, one of the first applications of Delphi, in 1948, was to use the "expert" judgements or opinions of horse racing handicappers to obtain a better estimate of a horse's chance of winning (Quade, 1968, p 334). Since then it has been used by several different organizations in a wide variety of applications. For example, Corning Glass Works used the Delphi technique to forecast electronic sales in the consumer oriented, industrial, and governmental business sectors five and ten years in the future (Johnson, 1976, pp 52-56). The U.S. Air Force used this technique in an attempt to quantify the distribution of quality required by the Air Force among the nonprior service accessions. It was felt that recruiting only enlisted personnel having a college degree (the high extreme) would lead to inefficiency and boredom, while recruiting only non-high school graduates would not permit the Air Force to carry out its mission. The Delphi technique was applied in an attempt to quantify a distribution of quality that would enable the Air Force to do its job (Taylor, et al, 1972, p 44).

#### 3. How A Delphi Experiment Is Conducted.

With this brief overview of what the Delphi technique is, it is appropriate to explain how a Delphi experiment would be conducted. Of course, the problem to be answered must have been determined and the decision to use the Delphi technique made in advance.

The first step would be to select the panel of "experts", personnel with some knowledge and/or experience in the issues to be addressed. But how many experts are needed? What background qualifies a person as an "expert"?

The first question is easier to answer. Research has shown that 10 to 15 panel members are generally sufficient to furnish reliable results. Fewer than 10 members may not provide adequate information and feedback to obtain reliable results, while more than 15 may seriously complicate the handling of the data (Johnson, 1976, p 52).

The second question, "What constitutes an expert?" is more difficult. One method of side-stepping the answer is to evaluate, or rank order, the panel members by their demonstrated "expertise" in the field, and then assign relative weights to their inputs after the responses are in (Helmer, 1963, p 5). A method that might be used to rank order the panel members--to be able to say that Mr. X is more of an "expert" than Mr. Y--would be to have potential panel members

name the person he believes is the most knowledgeable in the field. By determining those who received the most votes, a panel of highly qualified experts could be selected (Nat Def Uni, 1976, p 2). Another method might be for those conducting the Delphi experiment to rank order the potential panel members by their past performance: How correct have they been in previous forecasts.

However, Olaf Helmer's original Delphi applications did not consider the relative degree of expertise of the panel. What is required is that the viewpoints of all panel members have a chance of being heard.

Another important point concerning the panel and the Delphi technique is that of nonattribution, or anonymity. Panel members must be free to state the reasons for their beliefs or choices without fear of ridicule (Quade, 1968, p 334).

A final question concerning panel membership--Are there enough resources within the organization from which a panel of experts can be drawn, or must we go outside the organization; and are these experts willing to participate? The willingness to participate factor can be enhanced if, in the beginning of the experiment, the benefits and rewards to be gained by the panel members be explained. To accomplish this, three factors should be adequately covered at the start:

1. The purpose of the study,
2. An explanation of the Delphi technique and why it was selected, and
3. The benefits to the panel members of participation in the experiment.

As Jeffrey L. Johnson pointed out (Johnson, 1976, p 53), the Delphi technique is based upon feedback. The panel members have the benefit of the opinions of other experts and will be increasing their own knowledge. If they are unfamiliar with the Delphi technique, they will be able to learn it, and will possibly be able to apply it in doing their own forecasting in the future. These benefits, by themselves, may be sufficient to encourage enough participation from within the organization. If not, the only recourse available are to go outside the organization, and, possibly, have to pay for the experts' knowledge and opinions.

#### 4. A Delphi Example.

A graduate meteorological course in applied statistics given by St. Louis University at Scott Air Force Base, Illinois in the spring of 1977, offered an excellent opportunity to conduct a Delphi experiment. Most of the students in the class were Air Force weather officers, or officers who had had considerable experience in the weather career field. In addition, there were other personnel available at Scott AFB working in this field, e.g., the Weather Service War Planner in the headquarters of the Military Airlift Command.

As part of the class project, it was decided that by conducting a Delphi experiment, the class members, by serving on the panel, would gain some understanding of the Delphi procedures and, at the same time, it might be possible to answer a difficult meteorological question--namely, "What factors - if any - are causing the weather to change?"

The topic is a broad one; therefore, it was necessary in the beginning to restrict the area of discussion. By weather was meant, for example, last winter's extreme cold spell in the midwest and the drought in the Pacific northwest. By change was meant a trend over a period of five to ten years (as opposed to day to day changes or changes occurring over thousands of years.) So what we were interested in discovering were:

1. Is the weather changing?
2. If it is changing, what factors could be causing these changes? and
3. What is the relative importance of each of these factors in affecting a change in the weather?

With this overview of the purpose of the study, the Delphi technique was then explained to the class. They would be asked for their opinions on what was causing the weather to change. These opinions would then be collected, analyzed, and returned to them for a next iteration. In this way, everyone would have the benefit of the entire class' knowledge, and would be able to evaluate their own responses in this light.

The next several pages are the questionnaires which were submitted to the class. The first questionnaire on page 9-3 served to elicit fresh ideas on what factors may contribute to affecting a change in the weather. Thirty-four different factors were listed by the class in response to that questionnaire. These are listed in the second iteration questionnaire on pages 9-4 and 9-5. Of particular note is item number 34 on page 9-5. Well over half of the class indicated that there had been no significant change in the weather; that the unusual cold or drought spells that have been experienced are well within accepted probability limits. However, at the same time, those who indicated that there had been no significant change in the weather also listed other factors (from the first 33 on the list) that were causing the weather to change.

Therefore, in order to continue with the experiment, to more fully demonstrate the Delphi technique, item 34 was omitted from any further consideration. It would have been fruitless to ask what is causing the weather to change if all agree that the weather is not changing. This could be the real conclusion of the study. The variability of the weather that has been experienced in the last few years may well be just normal variation.

However, to continue the Delphi experiment further, it was decided to delete this response.

## WHAT IS CAUSING THE WEATHER TO CHANGE?

### A DELPHI APPLICATION

This questionnaire has been developed to elicit your ideas and opinions on what factors have an influence on changing the weather. This questionnaire is being used, rather than a face-to-face confrontation over a conference table, to enable you to express all of your ideas, even those you consider half-baked or far out. In other words, it's a brainstorming session without fear of ridicule and an iterative process with feedback.

We have two primary objectives in this effort. The first will be to identify those factors that may be causing a change in the weather (for example, exhausts from the internal combustion engine); while the second objective will be to measure or quantify the contribution of each factor in changing the weather. (For example, industrialization of agriculture--permitting vast areas of land to be denuded of vegetation--may be a more significant factor than automobile emissions in affecting the weather.)

With this as an overview, we request that you list in the space below those factors that you believe may be influencing the weather. The results will be analyzed, tabulated and returned to you during the next class meeting.



# WHAT IS CAUSING THE WEATHER TO CHANGE? A DELPHI APPLICATION

2. The following tabulation represents those factors you have indicated which may be influencing the weather. We would like your judgement of the relative importance of each of these factors. Specifically--in the short range of five to ten years--how important was each factor in affecting a change in the weather? For each factor, please check only one box, consider each factor separately.

FACTOR	VERY IMPORTANT	IMPORTANT	SLIGHTLY IMPORTANT	UNIMPORTANT
1. Increase in carbon dioxide in upper atmosphere (products of combustion).				
2. Solar cycles causing change in average earth temperature.				
3. Changes in earth's internal magnetic field.				
4. Changes in ocean currents.				
5. Heat island effect of large cities.				
6. Previous year's large deviations from mean climatic conditions.				
7. Ionospheric heating caused by radio energy propagation.				
8. Irrigation, water management activities.				
9. Geothermal heat source variations over large areas of land or beneath ocean floors.				
10. Increased agricultural efforts.				

FACTORS (Cont)	VERY IMPORTANT	IMPORTANT	SLIGHTLY IMPORTANT	UNIMPORTANT
11. Changes in ozone layer.				
12. Shift in long wave pattern that caused persisting high pressure ridge in Pacific/California region.				
13. Industrialization.				
14. Significant depreciation of total available potential energy contained on the earth.				
15. Rocket fuel affluents.				
16. Nuclear testing in the atmosphere.				
17. Underground nuclear tests.				
18. Radioactive isotopes in atmosphere.				
19. Change in position of polar jet stream.				
20. Pollution--causing increased number of condensation nuclei in the atmosphere.				
21. Changes in earth's albedo caused by irrigation; denuded of vegetation.				
22. Secret Soviet (Russian/Chinese) weapons.				
23. Increased airline flights.				
24. Satellites.				
25. Stratospheric warming/cooling.				
26. Change in infrared flux density in lower atmosphere.				
27. Increased coverage of land with concrete and asphalt, depleting aspiration sources of moisture.				
28. Vietnam War.				
29. Spray cans destroying the ozone layer.				
30. Weather modification efforts.				
31. Volcanic ash.				
32. Variation in the earth's elliptical orbit around the sun.				
33. Variation in the inclination of the earth's axis with respect to the elliptical orbit about the sun.				
34. No change (or change within normal variation).				

The second questionnaire was then given to the class. (At this point, item 34 was still being considered.) The class members--the Delphi panel--were then asked to evaluate the relative importance of each of the 34 factors in its ability to affect a change in the weather. They were asked to rate each factor either "Very Important," "Important," "Slightly Important," or "Unimportant." Table 1 summarizes the results. Two-thirds of the class (14 out of 21) believed that there has been no change in the weather, rating item 34 as either "Important" or "Very Important." On the other hand, one of the ideas submitted on the first iteration was rated "Unimportant" by the entire panel--number 24, satellites. This demonstrates one of the good features of Delphi--it allows all participants to express an idea, no matter how farfetched, without fear of ridicule.

DISTRIBUTION OF RESPONSES ON THE SECOND ITERATION QUESTIONNAIRE

Item No	Response	1	2	3	4
1		3	8	8	2
2		2	5	7	7
3		0	0	4	17
4		4	6	2	9
5		1	6	7	7
6		1	5	2	13
7		0	0	2	19
8		0	5	9	7
9		0	2	5	14
10		0	3	12	6
11		2	7	7	5
12		4	5	4	8
13		0	7	11	3
14		1	1	2	17
15		0	1	4	16
16		0	3	6	12
17		0	1	4	16
18		0	2	5	14
19		5	5	3	8
20		4	6	6	5
21		0	5	8	8
22		0	0	5	16
23		0	1	10	10
24		0	0	0	21
25		3	4	4	10
26		0	3	6	12
27		0	3	7	11
28		0	0	1	20
29		0	2	8	11
30		0	2	5	14
31		1	5	5	10
32		3	2	3	13
33		3	2	2	14
34		9	5	3	4

TABLE 1



The next problem was to evaluate these subjective ratings to determine which were the more important factors in influencing a change in the weather. Several weighting schemes were applied to the different subjective judgements. The following factors were the most important: pollution, a shift in the long wave pattern, change in the ozone layer, change in ocean currents, the heat island effect of large cities, changes in the earth's albedo, and geothermal heat source variations over a large area of the earth.

It was necessary to limit the number of important factors to be analyzed in the next iteration for two reasons. First, the technique described by Thomas L. Saaty (1973), using an eigenvalue analysis, was to be used to evaluate the relative importance or contribution of these factors. Unless the matrices developed are kept small, the task of evaluating or comparing each factor with every other factor, one at a time, becomes enormous. Second, the computer available to the author would solve the eigenvalue problem only for small (nine by nine) matrices. Therefore, only the eight factors mentioned above were included in the third iteration.

## WHAT IS CAUSING THE WEATHER TO CHANGE?

### A DELPHI APPLICATION

3. The attached tabulation represents those factors you have indicated which may be influencing the weather. We would like your judgement of the relative importance of each of these factors. Specifically--in the short range of five to ten years--how influential was each factor in affecting a change in the weather? As a preliminary means of establishing the relative importance of each factor, the following order is defined:

<u>DEGREE OF IMPORTANCE</u>	<u>DEFINITION</u>	<u>EXPLANATION</u>
1	EQUAL IMPORTANCE	TWO FACTORS ARE EQUAL IN THEIR ABILITY TO AFFECT WEATHER CHANGES
3	WEAK IMPORTANCE OF ONE FACTOR OVER ANOTHER	THERE IS SOME CON- VICTION THAT ONE FACTOR HAS MORE INFLUENCE THAN THE OTHER
5	STRONG IMPORTANCE OF ONE FACTOR OVER ANOTHER	STRONG BELIEF THAT LOGICAL CRITERIA EXIST TO SHOW THAT ONE FACTOR IS MORE IMPORTANT THAN THE OTHER FACTOR
7	DEMONSTRATED IMPOR- TANCE	ABSOLUTE CONVICTION AS TO THE IMPORTANCE OF ONE FACTOR OVER ANOTHER
2, 4, 6	INTERMEDIATE VALUES	WHEN COMPROMISE IS NEEDED
RECIPROCAL OF ABOVE NON-ZERO NUMBERS	IF FACTOR i HAS VALUE x WHEN COMPARED TO FACTOR j, THEN j HAS THE RECIPROCAL VALUE $\frac{1}{x}$ WHEN COMPARED WITH i x	

With these definitions, please evaluate each of the factors in the left column against the factors listed in the top row.

COMPARISON OF FACTORS WITH RESPECT  
TO INFLUENCE ON THE WEATHER

	POLLUTION OF THE ATMOSPHERE (MANY CAUSES)	SHIFT IN LONG WAVE PATTERN (AND/OR JET STREAM)	CHANGE IN OZONE LAYER	CHANGE IN OCEAN CURRENTS	SOLAR CYCLES	HEAT ISLAND EFFECT OF LARGE CITIES	CHANGES IN THE EARTH'S ALBEDO (INCLUDING AGRICULTURE)	GEO THERMAL HEAT SOURCE VARIATIONS OVER A LARGE AREA
POLLUTION OF THE ATMOSPHERE (MANY CAUSES)								
SHIFT IN LONG WAVE PATTERN (AND/OR JET STREAM)								
CHANGE IN OZONE LAYER								
CHANGE IN OCEAN CURRENTS								
SOLAR CYCLES								
HEAT ISLAND EFFECT OF LARGE CITIES								
CHANGES IN THE EARTH'S ALBEDO (INCLUDING AGRICULTURE)								
GEO THERMAL HEAT SOURCE VARIATIONS OVER A LARGE AREA								

The instructions for the third iteration questionnaire are shown on page 9-7. Each panel member was asked to compare each of the factors against every other factor, one at a time, and then to indicate the degree of influence one factor had over the other in affecting a change in the weather.

For example, items in the vertical column were compared against solar cycles across the top. For example, if it were believed that there existed logical criteria to show that pollution is more important than solar cycles in influencing the weather, the panel member would place a "5" in the first row, fifth column position. Then the reciprocal value "1/5" would be placed in the opposite position--in the fifth row, first column.

When two items were rated equal in their ability to influence the weather, or when a factor was rated against itself, a "1" would be placed in that position. Thus, the main diagonal of the matrix would contain only 1's.

The responses of all panel members were analyzed to obtain the median. With this information, the matrix shown on page 9-9 was obtained. The eigenvalue of this matrix is slightly less than 9, and, thus, not too far from the consistent value 8. The eigenvector corresponding to this largest eigenvalue, normalized so that the sum of the individual terms is 1, is given by:

$$X = (0.12 \ 0.18 \ 0.13 \ 0.10 \ 0.26 \ 0.06 \ 0.11 \ 0.04)$$

**COMPARISON OF FACTORS WITH RESPECT  
TO INFLUENCE ON THE WEATHER**

	POLLUTION OF THE ATMOSPHERE (MANY CAUSES)	SHIFT IN LONG WAVE PATTERN (AND/OR JET STREAM)	CHANGE IN OZONE LAYER	CHANGE IN OCEAN CURRENTS	SOLAR CYCLES	HEAT ISLAND EFFECT OF LARGE CITIES	CHANGES IN THE EARTH'S ALBEDO (INCLUDING AGRICULTURE)	GEO THERMAL HEAT SOURCE VARIATIONS OVER A LARGE AREA
POLLUTION OF THE ATMOSPHERE (MANY CAUSES)	1	1	1	1	1/5	2	3	3
SHIFT IN LONG WAVE PATTERN (AND/OR JET STREAM)	1	1	5	1	1	2	1	5
CHANGE IN OZONE LAYER	1	1/5	1	1	1	3	1	3
CHANGE IN OCEAN CURRENTS	1	1	1	1	1/3	2	1	3
SOLAR CYCLES	5	1	1	3	1	4	3	5
HEAT ISLAND EFFECT OF LARGE CITIES	1/2	1/2	1/3	1/2	1/4	1	1	1
CHANGES IN THE EARTH'S ALBEDO (INCLUDING AGRICULTURE)	1/3	1	1	1	1/3	1	1	3
GEO THERMAL HEAT SOURCE VARIATIONS OVER A LARGE AREA	1/3	1/5	1/3	1/3	1/5	1	1/3	1

The Delphi panel was then asked to reevaluate their previous inputs in light of this new information. The results of this fourth iteration did not change the matrix entries. Thus, the figures above are really the final result.

**5. Conclusion.**

The several conclusions resulting from this exercise fall into two areas.

First, with regard to the experiment itself, it must be emphasized how important it is that the instructions in the beginning be clear and understood by all of the participants. In this experiment, even to the last iteration, there were one or two panel members who were unclear about the purpose of the effort. Also, adequate time must be furnished to the participants to accurately complete the questionnaires. In this exercise, which was generally accomplished in class during a break period, not enough time was afforded the panel members to really evaluate their inputs, or to furnish sound reasons for their positions.





## CHAPTER 10

### RESULTS OF A SINGLE STATION FORECASTING EXPERIMENT

Robert G. Miller  
Roger C. Whiton  
Michael J. Kelly, Jr

Headquarters Air Weather Service  
Scott Air Force Base, Illinois

#### ABSTRACT

This paper describes a forecasting experiment performed using actual single station data for Rickenbacker AFB, Ohio. The model used, suggested by Miller (1968), employs the principle of multivariate regression in a Markov process. The model predicts the probability distribution of the following observed weather elements: temperature, dew-point depression, station pressure, crosswinds, wind direction and speed, sky cover, ceiling, visibility, and weather elements including haze and smoke, blowing conditions, fog and ground fog, freezing precipitation, rain and drizzle, snow, rain showers, snow showers, and also the condition of no weather. At forecast time the system uses all of these elements as predictors, plus day of year, time of day, and the most recent 3-hour pressure change. An extension is being formulated using eigenfunctions which allows predictions at any future time, not just at discrete points. Results of nearly 30,000 independent forecasts are described, where comparisons using the Brier score and hits are made against persistence and conditional climatology (ceiling and visibility only).

#### 1. INTRODUCTION

The Air Weather Service (AWS) requires a variety of approaches to weather forecasting in order to provide military weather support across a full spectrum of conflict scenarios. Peacetime weather support, strategic and tactical activities, and support to command and control agencies demand a highly developed, computer based, centralized production system. In certain tactical conflicts, on the other hand, there is a need for a stand-alone forecasting method, a technique that can be used skillfully by a weatherman, when circumstances deny his standard communications.

To meet the need for such a capability, we conducted a development effort, known as the Single Station Forecasting Experiment. This paper will cover the purpose of the experiment, the data used, the methodology employed, the results obtained, and possible extensions to the model.

#### 2. PURPOSE

There were at least six reasons for undertaking this experiment: (1) The AWS Commander requested it, (2) A single station capability was strongly advocated by the AWS wing commanders, (3) Military exercises suggested the desirability of a stand alone forecast capability, (4) Circumstances of no communication with the centralized production system required it, (5) A fast, compact simulator of both weather observations and forecasts was required for input into the Defense Department's war gaming models and (6) Automated statistical forecasting algorithms under development, such as found in the Modular Automated Weather System test at Scott AFB, could use multiple elements as predictors.

The capability would have value in two typical operational scenarios:

- New Forecaster - a relatively new, inexperienced forecaster deployed to Europe or Korea--first time in area, in need of guidance.
- Army Field Support - a non-communication situation. Must provide weather forecasts having only the current observation.

The goal was to provide the AWS forecaster with a stand alone capability for making probability forecasts.

### 3. DATA

Let us address the details of the experiment. First the data specifications:

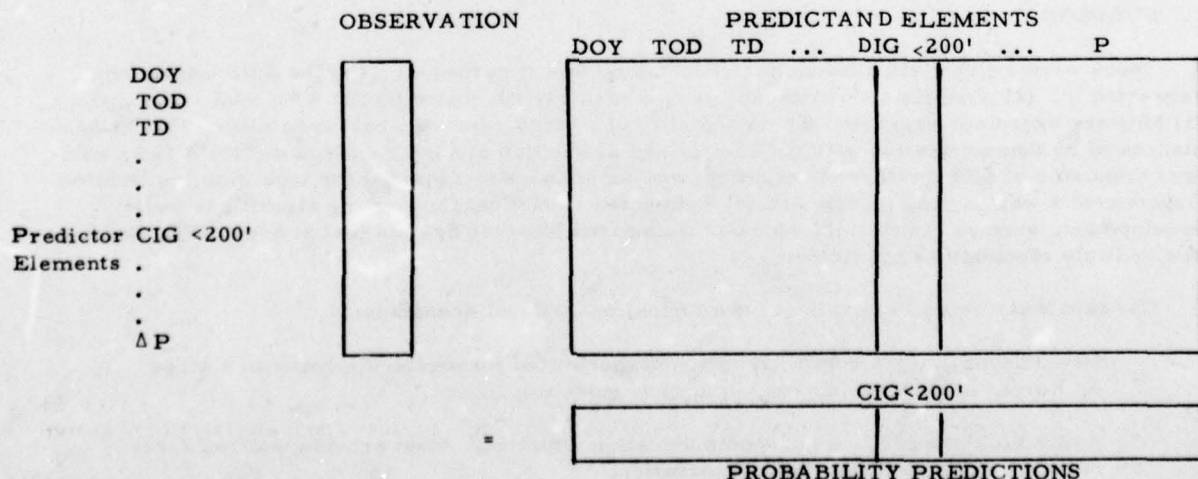
#### Specifications\*:

Location--Rickenbacker AFB, Ohio  
 Forecast intervals--3 hours  
 Predictors (single station only):  
   Day of year  
   Time of day  
   Temperature--dew-point depression  
   Sea level pressure  
   Cross wind  
   Wind rose  
   Wind speed  
   Temperature  
   Sky cover  
   Ceiling  
   Visibility  
   Weather (smoke, dust or haze; blowing snow, dust or sand; fog; frozen precipitation; rain or drizzle; snow; rain showers; snow showers; thunderstorms; none)  
   Pressure change (past 3 hours)  
 Elements predicted--same as above except DOY, TOD, which are deterministic and  $\Delta P$  which is irrelevant  
 Data sample--10 years dependent, 10 years independent (29,154 new forecast situations)

Note the absence of vital cloud information, cloud layer height, cloud layer amount, and cloud type for low, middle, or high conditions.

### 4. METHODOLOGY

The statistical method employed in this experiment is that of multivariate regression, suggested by Miller (1968). That is, many variables are predicted with the same set of predictors. One other condition also exists. All of the predictors are either one or zero (on or off). The regression predicts the probability that a condition is on or off. The following depicts such an arrangement.



\*See Appendix for details



The predictors are shown as rows, the predictands as columns. Note that both involve the exact same elements. For this problem there were 153 predictors.

The objective is to multiply the observation, made up of ones and zeros, times the regression coefficients in a particular column, such as the one shown for ceiling <200'. The result is the probability that in three hours, a ceiling <200' will be observed. That probability is entered into the probability prediction row, as are all the other predicted probabilities.

Mathematically this is expressed as

$$\underline{\hat{P}} = \underline{\theta}' \underline{A}$$

where  $\underline{\hat{P}}$  is the row vector of predicted probabilities,  $\underline{\theta}'$  is the transpose of the observation column vector  $\underline{\theta}$ , and  $\underline{A}$  is the matrix of regression coefficients where each column constitutes the regression coefficients for a particular predictand. This formulation generalizes to

$$\underline{\hat{P}}_T = \underline{\theta}' \underline{A}^T$$

when an estimate of the probability is desired, under a Markov assumption, for T units of time into the future. The basic unit of time in this experiment was 3 hours. Therefore, to estimate the probabilities for 6 hours, T=2.

## 5. RESULTS

The following summarizes the results for independent 3 and 6 hour forecasts on 29, 154 forecast situations:

- Regression is superior to persistence on number of hits for all variables except snow at 3 hours.
- Regression is superior to conditional climatology in terms of hits and Brier score for visibility and ceiling.

Specifically, for 3 hour forecasts, the comparison between multivariate regression and persistence is shown below in terms of the number of correct forecasts (hits):

Weather Elements	NUMBER OF HITS		
	Regression	Persistence	Difference
Wind speed	28022	27761	+ 261
Cross wind	28932	28848	+ 84
Temperature	22299	22152	+ 147
Visibility	24061	23317	+ 744
Ceiling	21360	20995	+ 365
Sky cover	17361	17313	+ 48
Rain	27990	27944	+ 46
Rain showers	28327	27902	+ 425
Snow	28427	28496	- 69
Snow showers	28779	28616	+ 163
Thunderstorms	28894	28729	+ 165
Freezing precipitation	29114	29096	+ 18
Smoke or haze	25292	25026	+ 266
Blowing snow or sand	29100	29090	+ 10
No weather	24063	23776	+ 287

In making this comparison, the probabilistic regression forecasts were categorized by selecting the condition with the highest forecast probability. Remember that persistence is a formidable competitor at 3 hours. The visibility and ceiling results are especially impressive.

The 6 hour comparisons are as follows:

Weather Elements	NUMBER OF HITS		
	Regression	Persistence	Difference
Wind speed	27984	27390	+ 594
Cross wind	28928	28766	+ 162
Temperature	19152	17253	+ 1899
Visibility	23439	21807	+ 1632
Ceiling	19423	18300	+ 1123
Sky cover	15066	14162	+ 904
Rain	27870	27528	+ 342
Rain showers	28323	27706	+ 617
Snow	28374	28238	+ 136
Snow showers	28775	28544	+ 231
Thunderstorms	28890	28656	+ 234
Freezing precipitation	29110	29076	+ 34
Smoke or haze	23431	23104	+ 327
Blowing snow or sand	29096	29084	+ 12
No weather	22344	21664	+ 680

For verifying the probabilities, regression is compared with the stiff competitor, conditional climatology, using the well known Brier score. Smaller values are better. The differences are highly significant statistically. That is, the likelihood of achieving differences of these magnitudes by pure chance is practically zero.

#### COMPARATIVE STATISTIC Brier Probability Score\*

Weather Element		Regression	Conditional Climatology	Difference
3 hr	Visibility	.2564	.2732	.0169**
	Ceiling	.3755	.4043	.0288**
6 hr	Visibility	.2998	.3175	.0177**
	Ceiling	.4397	.4763	.0366**

#### 6. DISCUSSION

The statistical method employed does not require a large computer to make the forecasts. It did require some heavy computing to achieve the results, for which we acknowledge the assistance of these installations: Saint Louis University, the USAF Environmental Technical Applications Center, the Military Airlift Command Data Automation, and the Defense Commercial Communications Office.

The more important features of the procedure are:

Skillful - Consistently superior to persistence and conditional climatology.

Objective - No judgment is needed. Two different people should get the identical answer.

Distribution Free - No need to make any assumption such as normality.

\* Smaller values are better

\*\* Highly statistically significant

Multivariate - Same predictors used for many predictands.

Fast - Operates in real time.

Easy to Use - Can be run on a small calculator or by hand.

Asynoptic Data - Should a special observation be warranted, a new prediction can be made with these data. PIREPs, radar, satellite, or intermediate times are no real problem.

Nonlinear - The zero-one predictors are weighted separately over the range of the original weather variable.

Interpretable - To see the effect of any observed condition on the predicted probability, just look at its coefficient.

Variable Threshold - It is a simple matter to set new limits on the predictand desired. Merely add the coefficients or the probabilities for the categories needed. Since all weather elements have only a finite number of digits of accuracy, it will require only the inclusion of each possible value as a predictand.

Growth Potential - Improvements in statistical methods, computers, observational accuracy, or new measurements, like satellite, radar, solar, can be added to make the forecasts better. Incidentally, other things might also enhance the forecast accuracy such as: one hour equations, stepwise selection of predictors, time-change predictors, and interactions among predictors.

## 7. EXTENSIONS OF THE MODEL

To make predictions at 6 hours, the 3 hour regression coefficient matrix was squared utilizing a Markov assumption. Another approach to making a 6 hour forecast from the 3 hour regression coefficients would have been to enter the 3 hour forecasted probabilities into the observation vector and have it reprocessed. A more general alternative would be to use eigenfunctions. This would, in principal, permit predictions into the future with time as a continuum.

The formulation of this latter alternative has been made and tested on a more limited set of data than was used in the single station experiment. It worked as expected. Furthermore, the need for computer storage was greatly reduced, since eigenvectors were used instead of a coefficient matrix.

When the predictors include time of day and day of year, the eigensolution produces complex roots and vectors. A computational difficulty arises when solving large matrices of this type. Research on this problem is in progress.

Another extension in the model introduces the concept of generalized operators, where one matrix of regression coefficients applies at more than one geographical location. Similar applications have been successful in other contexts (see Harris, Bryan and MacMonegle, 1963 and 1965). Research in this area is continuing.



## Appendix A: Details of the Single Station Forecast Experiment

This Appendix describes the data used and the predictor/predictand categories for the experiment.

A magnetic tape of hourly surface weather observations for Rickenbacker AFB, Ohio, was obtained from the United States Air Force Environmental Technical Applications Center. These data were originally hand punched and are believed to be of good quality. The expected quality of the data, and time constraints dictated that we not edit the data. The dependent (1946-1955) and independent (1956-1965) samples consisted of 24,384 and 29,154 observations respectively. If an observation contained a missing element, it was not used.

Table A-1 describes the elements and categories that were verified. They were selected on the basis of operational significance. For example, read the three wind speed categories as 1) 0 to  $\leq 14$ , 2) greater than 14 but  $\leq 24$  and 3) greater than 24.

Table A-2 describes the predictor elements and categories. The predictors were chosen subjectively. Elements were chosen because they were available on the data tape. Cloud type, for example, was not used as a predictor, since it was not available for part of the 20-year period of record. Predictor categories were defined to align with the verification categories, and to assure an adequate number of occurrences for each category.

Table A-1. Verification Elements and Categories

<u>Element</u>	<u>No. of Categories</u>	<u>Categories</u>
Wind Speed (Kts)	3	$0 \leq 14 \leq 24 < \infty$
Cross Wind (Kts)	2	$0 \leq 14 < \infty$
Temperature ( $^{\circ}\text{F}$ )	6	Absolute $0 \leq 15 \leq 31 \leq 49 \leq 67 \leq 84 < \infty$
Visibility (mi)	6	$0 < \frac{1}{2} < 1 < 2 < 3 < 6 < \infty$
Ceiling Height (ft)	6	$0 < 200 < 500 < 1000 < 3000 < 10000 < \infty$
Sky Cover	4	Clear or partial obscuration; scattered; broken; overcast or total obscuration
Rain	2	Yes or No
Rain Showers	2	Yes or No
Snow	2	Yes or No
Snow Showers	2	Yes or No
Thunderstorm	2	Yes or No
Freezing	2	Yes or No
Fog, Haze, or Smoke	2	Yes or No
Blowing Dust, Sand, or Snow	2	Yes or No
No Weather	2	Yes or No

Table A-2. Predictor Elements and Categories

<u>Element</u>	<u>No. of Categories</u>	<u>Categories</u>
Day of Year	26	$1 \leq 14 \leq 28 \leq \dots \leq 366$
Hour of Day (Z)	8	00; 03; 06; 09; 12; 15; 18; 21
Temperature-Dewpoint Depression ( $^{\circ}$ )	11	$0 < 1 < 2 < 3 < 4 < 5 < 6 < 8 < 10 < 15 < 20 < \infty$
Sea-Level Pressure (mb)	5	$0 \leq 1000 \leq 1010 \leq 1020 \leq 1035 < \infty$
Crosswind (Kts)	2	$0 \leq 14 \leq \infty$
Wind Rose ( $^{\circ}$ and Kts)	17	Calm and variable; 1- 45, $0 < 10 < \infty$ 46- 90, $0 < 10 < \infty$ 91-135, $0 < 10 < \infty$ 136-180, $0 < 10 < \infty$ 181-225, $0 < 15 < \infty$ 226-270, $0 < 15 < \infty$ 271-315, $0 < 15 < \infty$ 316-360, $0 < 15 < \infty$
Wind Speed (Kts)	6	$0 < 2 < 6 < 10 < 15 < 20 < \infty$
Temperature ( $^{\circ}$ F)	13	Absolute $0 \leq 5 \leq 10 \leq 15 \leq 24 \leq 31 \leq 40 \leq 49 \leq 58$ $\leq 67 \leq 76 \leq 84 \leq 89 < \infty$
Sky Cover	4	Clear or partial obscuration; scattered; broken; overcast or total obscuration
Ceiling Height (ft)	21	$0 < 200 < 400 < 500 < 600 < 700 < 800 < 900$ $< 1000 < 1100 < 1600 < 2100 < 2600 < 3000$ $< 3600 < 4100 < 6000 < 10000 < 11000 < 13000$ < unlimited; unlimited
Visibility (mi)	12	$0 < \frac{1}{2} < \frac{3}{4} < 1 < 1\frac{1}{2} < 2 < 2\frac{1}{2} < 3 < 4 < 5 < 6$ $< 7 < \infty$
Smoke, Haze, Dust	2	Yes or No
Blowing Dust, Sand, or Snow	2	Yes or No
Fog	2	Yes or No
Rain or Drizzle	2	Yes or No
Freezing Rain or Drizzle	2	Yes or No
Rain	2	Yes or No
Snow	2	Yes or No
Rain Showers	2	Yes or No
Snow Showers	2	Yes or No
Thunderstorm	2	Yes or No
No Weather	2	Yes or No
Pressure Change (3 hour, mb)	7	$-\infty < -3.9 < -2.0 < -.9 < 1.0 < 2.0 < 4.0 < \infty$

# BIBLIOGRAPHY

- Alaka, M. A., W. D. Bonner, J. P. Charba, R. L. Crisci, R. C. Elvander, and R. M. Reap. "Objective Techniques for Forecasting Thunderstorms and Severe Weather." Department of Transportation Report, No FAA-RD-73-117, Washington, D.C., FAA, 1973, 97 pp.
- Anderson, T. W. An Introduction to Multivariate Analysis. John Wiley and Sons, Inc, New York, 1958.
- AWSR 105-13. "Probability Forecasts and Mission Success Indicators." Air Weather Service, USAF, Scott AFB, Illinois, 17 January 1977, 4 pp.
- Ayers, F., Jr. Theory and Problems of Matrices. Schaum Publishing Co, New York, 1962, 219 pp.
- Bartlett, M. S. "Square Root Transformation in Analysis of Variance." Supplement to the Journal of the Royal Statistical Society, Vol 3, 1936, pp 74-75.
- Bartlett, M. S. "Some Examples of Statistical Methods of Research in Agriculture and Applied Biology." Supplement to the Journal of the Royal Statistical Society, Vol 4, 1937, p 168.
- Breiman, L. Probability and Stochastic Processes: With a View Toward Applications. Houghton Mifflin Co, Boston, 1969, pp 152-216.
- Brier, Glenn W. and Roger A. Allen. "Verification of Weather Forecasts." Compendium of Meteorology, Ed, T. F. Malone. American Meteorological Society, Boston, 1951, pp 841-848.
- Bross, Irwin D. J. Design for Decision. The Macmillan Co, New York, 1953, 276 pp.
- Bryan, Joseph G., and Isadore Enger. "Use of Probability Forecasts to Maximize Various Skill Scores." Journal of Applied Meteorology, Vol 6, No 5, 1967, pp 762-769.
- Bryan, J. G. and A. Singer. "Prediction of Reenlistment Using Regression Estimation of Event Probabilities (REEP)." INS Research Contribution No 13, 1 October 1965.
- Bryan, J. G. and J. R. Southan. Optimum Subdivision of a Variable by the Method of D. R. Cox. Report No TRC-21, Contract No AF19(604) 5207, The Travelers Research Center, Inc, Hartford, 1962.
- Bryan, J. G. "M331 Notes." Unpublished.
- Bryan, Joseph G. "The Generalized Discriminant Function: Mathematical Foundation and Computational Routine." Harvard Educational Review, Vol XXI, No 2, 1951, pp 90-95.
- Caskey, J. E. "A Markov Chain Model for the Probability of Precipitation Occurrence in Intervals of Various Length." Monthly Weather Review, 91, 1963, pp 298-301.
- Caskey, J. E. "Markov Chain Model of Cold Spells at London." Met Magazine, 93, 1964, pp 136-138.
- Charba, J. P. "Operational Scheme for Short Range Forecasts of Severe Local Weather." Preprints of Ninth Conference on Severe Local Storms, American Meteorological Society, Boston, MA, 1975, pp 51-57.
- Chisholm, D. A. "Objective Prediction of Mesoscale Variations of Sensor Equivalent Visibility During Advective Situations." AFGL-TR-76-0132, Environmental Research Papers No 569, Air Force Geophysics Laboratory, 1976, 31 pp.
- Cooley, W. W. and P. R. Lohnes. Multivariate Procedures for the Behavioral Sciences, John Wiley and Sons, Inc, New York, 1962.
- Cox, D. R. "Note on Grouping." Journal of the American Statistical Assoc, Vol 52, No 280, December 1957, pp 543-547.
- Cox, D. R., Some Procedures Connected With the Logistic Qualitative Response Curve, in Research Papers in Statistics, a Festschrift for J. Neyman, edited by F. N. David, John Wiley and Sons, New York, 1966, pp. 55-71.



- Crout, P. D. "A Short Method of Evaluating Determinants and Solving Systems of Linear Equations." Trans AIEE, 60: 1235, 1941.
- David, C. L. "An Objective Method for Estimating the Probability of Severe Thunderstorms Using Predictors from the NMC (PE) Numerical Prediction Model and from Observed Surface Data." Preprints of 8th Conference on Severe Local Storms, American Meteorological Society, Boston, MA, 1973, pp 223-225.
- Draper, N. R. and H. Smith. Applied Regression Analysis. John Wiley and Sons, Inc, New York, 407 pp.
- Eichmeier, A. H. and W. D. Baten. "Rainfall Probabilities During the Crop Season in Southern Lower Michigan." Monthly Weather Review, 91, 1962, pp 298-301.
- Ezekiel, M. and K. A. Fox. Methods of Correlation and Regression Analysis. John Wiley and Sons, Inc, New York, 1959, 548 pp.
- Feller, W. An Introduction to Probability Theory and Its Applications, Vol I, 1st Ed. John Wiley and Sons, Inc, New York, 1950, p 217.
- Feller, W. An Introduction to Probability Theory and Its Applications, Vol I, 3rd Ed, Rev. John Wiley and Sons, Inc, New York, 1968, pp 372-427.
- Feller, W. An Introduction to Probability Theory and Its Applications, Vol II. John Wiley and Sons, Inc, New York, 1966, 626 pp.
- Feyerherm, A. M. and L. D. Bark. "Statistical Methods for Persistent Precipitation Patterns." Journal of Applied Meteorology, 4, 1965, pp 320-328.
- Fisher, R. A. "On the Dominance Ratio." Proceedings of the Royal Society of Edenberg, Vol 42, 1921-1922, p 326.
- Fix, Evelyn and Joseph L. Hodges, Jr. Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties. Report No 4, USAF School of Aviation Medicine, Randolph Field, TX, 1951.
- Flegg, H. Graham. Boolean Algebra and Its Application. John Wiley and Sons, New York, 1964.
- Freeman, M. F. and J. W. Tukey. "Transformations Related to the Angular and the Square Root." Annals of Mathematical Statistics, Vol 21, 1950.
- Frye, Alton. "US Space Policy: An Example of Political Analysis." Systems Analysis and Policy Planning: Applications in Defense, American Elsevier Publishing Co, Inc, New York, 1968.
- Gabriel, K. R. "The Distribution of the Number of Successes in a Sequence of Dependent Trials." Biometrika, 46, 1959, pp 454-460.
- Gabriel, K. R. and J. Neumann. "On the Distribution of Weather Cycles by Length." Q. J. Roy. Met Soc, 83, 1957, pp 375-380.
- Gabriel, K. R. and J. Neumann. "A Markov Chain Model for Daily Rainfall Occurrence at Tel Aviv." Q. J. Roy. Met. Soc, 88, 1962, pp 90-95.
- George, J. J. Weather Forecasting for Aeronautics. Academic Press, New York, 1960, pp 540-650.
- Glahn, H. R. and D. A. Lowry. "The Use of Model Output Statistics (MOS) in Objective Weather Forecasting." Journal of Applied Meteorology, Amer Meteorological Society, Boston, MA, 11, 1972, pp 1203-1211.
- Gringorten, I. I. "A Stochastic Model of the Frequency and Duration of Weather Events." Journal of Applied Meteorology, 5, 1966, pp 606-624.
- Gringorten, Irving I. "Verification to Determine and Measure Forecasting Skill." Journal of Applied Meteorology, Vol 6, No 5, 1967, pp 742-747.
- Gringorten, I. I. "Estimating Finite-Time Maxima and Minima of a Stationary Gaussian Ornstein-Uhlenbeck Process by Monte Carlo Simulation." American Stat Assn Journal, 63, 1968, pp 1517-1521.

- Gringorten, I. I. "Modelling Conditional Probability." Journal of Applied Meteorology, 10, 1971, pp 646-657.
- Gringorten, I. I. "Conditional Probability for an Exact, Non-Categorized Initial Condition." Monthly Weather Review, 100, 1972, pp 796-798.
- Harris, R. J., J. G. Bryan, and J. E. MacMonegle. Terminal Weather Prediction Studies, Technical Note No 3, Contract AF19(626)-16, System 433L, The Travelers Research Center, Inc, 1963, 264 pp.
- Harris, R. G., J. G. Bryan, and J. E. MacMonegle. Diagnosis of Surface Weather Conditions from Observed and Prognostic Upper Air Parameters, 433L Systems Program Office, Electronic Systems Div, Air Force Systems Command, ESD-TR-65-2, 1965, 135 pp.
- Heidke, P., "Berechnung des Erfolges und der Gute der Windstarkevorhersagen im Sturmwarnungsdienst," Geografische Annaler, Stockholm, 1926, Vol 8, pp 310-349.
- Heinrich, Barbara Ann. Expert Opinion About Uncertainty. Scientific Document N00014-67-A-0103-0011 prepared for the Personnel and Training Research Programs Office, Office of Naval Research, April, 1971.
- Helmer, Olaf. The Systematic Use of Expert Judgment in Operations Research. Paper prepared for presentation at the Third International Conference on Operational Research, Oslo, Norway, July, 1963.
- Hering, W. S. Personal communication, 1977.
- Hering, W. S. and D. L. Quick. "Hanscom Visibility Forecasting Experiments." Preprints, 5th Conf on Weather Forecasting and Analysis, American Meteorological Society, 1974, pp 224-227.
- Hildebrand, F. B. "Introduction to Numerical Analysis." McGraw-Hill Book Co, Inc, 1956.
- Hotelling, H. "The Most Predictable Criterion." Journal of Educational Psychology 26, 1935, pp 139-142.
- Hotelling, H. "Relations Between Two Sets of Variates." Biometrika 28, pp 321-377.
- Howard, R. A. Dynamic Programming and Markov Processes. Technology Press of the Massachusetts Institute of Technology, Boston, and John Wiley & Sons, Inc, New York, 1960, 136 pp.
- Hughes, L. A., F. Baer, G. E. Birchfield, and R. E. Kaylor. "Hurricane Hazel and a Long-Wave Outlook." Bulletin of the American Meteorological Society, Vol 36, 1955, pp 528-533.
- Johnson, Jeffry L. "A Ten-Year Delphi Forecast in the Electronics Industry." Management Review, Vol 65, August 1976, pp 52-56.
- Kenney, J. F. and E. S. Keeping. Mathematics of Statistics, Part 2, Second Edition. D. Van Nostrand Co, New York, 1951.
- Knox, J. L. "The Storm 'Hazel'." Bulletin of the American Meteorological Society, Vol 36, 1955, pp 239-246.
- Kunz, K. S. "Numerical Analysis." McGraw-Hill Book Co, Inc, 1957.
- Lowry, W. P. and D. Guthrie. "Markov Chains of Order Greater than One." Monthly Weather Review, 96, 1968, pp 798-801.
- Malone, T. F. "Studies in Statistical Weather Prediction." Final Report, Travelers Weather Research Center, The Travelers Insurance Co, Hartford, Conn, 1958, 238 pp.
- Miller, R. C. "Notes on Analysis and Severe Forecasting Procedures of the Air Force Global Weather Central." Technical Report 200 (Rev), AWS, USAF, 1972.
- Miller, R. G. Statistics and the predictability of weather in Studies in Statistical Weather Prediction, Final Report, AF19(604)-1590, Travelers Weather Research Center, 1958, pp 137-153.
- Miller, R. G. The screening procedure in Studies in Statistical Weather Prediction, Final Report, AF19(604)-1590, Travelers Weather Research Center, 1958, pp 86-96.
- Miller, R. G., An Application of Multiple Discriminant Analysis to the Probabilistic Prediction of Meteorological Conditions Affecting Operational Decisions. Technical Memorandum No 4, The Travelers Research Center, Inc, Hartford, Conn, 1961.

- Miller, R. G., "Statistical Prediction by Discriminant Analysis." Meteorological Monographs, Vol 4, No 25, American Meteorological Society, Boston, MA, 1962, 54 pp.
- Miller, R. G., "Regression Estimation of Event Probabilities." Tech Report 7411-121, Contract Cwb-10704, The Travelers Research Center, Inc, Hartford, Conn, 1964.
- Miller, R. G., "A Stochastic Model for Real-Time On-Demand Weather Predictions." Proceedings, 1st Statistical Meteorological Conference, American Meteorological Society, 1968, pp 48-51.
- Miller, R. G., Final Report Under Letter Contract Dated April 8, 1968. Report prepared for F. M. D. Richardson, Acting Director, Bureau of Medical Review, August, 1968.
- Miller, R. G., SLAM: A Screening Lattice Algorithm for Non-linear Regression Estimation of Event Probabilities, The Travelers Research Center, Inc, Hartford, Conn, 1969, 23pp.
- Miller, R. G., Precis of Inflation, Technology and Growth, Possible Long-Range Implications for Insurance, July, 1973.
- Miller, R. G. "A Technical Description of LIMRA's Company Buyer Service." Life Insurance Marketing and Research Association, June, 1976.
- Miller, R. G. "Canonical Correlation Applied to Life Insurance Market Research." Life Insurance Marketing and Research Association, 1976.
- Mizrahi, A. and M. Sullivan. Finite Mathematics with Applications: For Business and Social Sciences. John Wiley & Sons, Inc, New York, 1973, pp 313-339.
- Murphy, Allan H. and Edward S. Epstein. "Verification of Probabilistic Predictions: A Brief Review." Journal of Applied Meteorology, Vol 6, No 5, 1967, pp 748-755.
- Murphy, A. H. and R. A. Allen. "Probabilistic Prediction in Meteorology: A Bibliography." ESSA Technical Memorandum WBTM TDL 35, Silver Spring, MD, 1970, 60 pp.
- National Defense University, Study on Climatology, undated, circa 1976.
- Quade, E. S. "When Quantitative Models are Inadequate." Systems Analysis and Policy Planning: Applications in Defense, American Elsevier Publishing Co, New York, 1968.
- Rao, C., Radhakrishna. Advanced Statistical Methods in Biometric Research. John Wiley & Sons, 1952, pp 257-258.
- Rapp, R. R. and A. H. Isnardi. Variability of Upper Winds. Progress Rep 138-03, Signal Corps Contract No DA36-039 SC-72, New York University, New York, 1951.
- Reap, R. M. "Thunderstorm and Severe Weather Probabilities Based on Model Output Statistics." Preprints of Fifth Conf on Weather Forecasting and Analysis, Amer Meteor Soc, Boston, MA, 1974, pp 266-269.
- Reap, R. M. and D. S. Foster. "New Operational Thunderstorm and Severe Storm Probability Forecasts Based on Model Output Statistics (MOS)." Preprints of the Ninth Conf on Severe Local Storms, American Meteorological Society, Boston, MA, 1975, pp 58-63.
- Rulon, Phillip J. "Distinctions Between Discriminant and Regression Analyses and a Geometric Interpretation of the Discriminant Function." Harvard Educational Review, Vol XXI, No 2, 1951, pp 60-90.
- Saaty, Thomas L. "Hierarchies and Priorities - Eigenvalue Analysis." Mimeographed, 1973.
- Saaty, Thomas L. and Mohamad W. Khouja. "A Measure of World Influence." Journal of Peace Science, spring, 1976, pp 31-48.
- Sakamoto, C. M. "Markov Chain Models for Probabilities of Hot and Cool Days Sequences and Hot Spells in Nevada." ESSA Technical Report EDS-9, ESSA Environmental Data Service, Silver Spring, 1970, 26 pp.
- Shannon, Claude E. and Warren Weaver. Mathematical Theory of Communication. Urbana, University of Illinois Press, 1949, 117 pp.



- Siegel, S., Nonparametric Statistics for the Behavioral Sciences, New York, McGraw-Hill Book Co., 1956, 312 pp.
- Snedecor, G. W., "Statistical Methods" Iowa State Press, 1946.
- Sorenson, E. L., Isadore Enger, and Thomas G. Johnson, "Users for Statistical Computer Program Package." Travelers Research Center, Report 7463-158, 1965.
- Spiegel, H. J., "An Investigation of Three Markovian Chain Models for Determination of Probabilities of Hot and Cool Spells." MS Thesis, Rutgers, 1966, 95 pp.
- Tahnk, W. R., "Objective Prediction of Fine Scale Variations in Radiation Fog Intensity." AFCRL-TR-75-0269, AF Surveys in Geophysics No 311, Air Force Cambridge Research Lab, 1975, 37 pp.
- Tanur, J. M., et al, Statistics: A Guide to the Unknown. Holden-Day, Inc, San Francisco, 1972.
- Tatsuoka, M. M., Multivariate Analysis: Techniques for Educational and Psychological Research. John Wiley & Sons, New York, 1971.
- Taylor, J. W., Leonard V. Scifers, and William J. Janeczek, Palace Quality. Report AF/DPXY-PR-72-006, USAF, Directorate of Personnel Plans, May, 1972.
- Tiedeman, D. V., "The Utility of the Discriminant Function in Psychological and Guidance Investigations." Harvard Educational Review, Vol XXI, No 2, pp 71-79.
- Tukey, J. W., Comparing Individual Means in the Analysis of Variance, Biometrics, 1949, Vol 5, pp 99-114.
- Tukey, J. W., "On the Comparative Anatomy of Transformations." Annals of Mathematical Statistics, Vol 28, No 3, September 1957.
- Walker, H. M. and J. Lev., Statistical Inference, New York, Holt, Reinhart, and Winston, 1953.
- Walker, H. M. and J. Lev., Elementary Statistical Methods. Henry Holt and Co, New York, 1958, 302 pp.
- Yamane, T., Statistics: An Introductory Analysis. Harper and Row, New York, 1967.

# Appendix A

## PLODITE PROGRAM

```

1  CPLDT PLODITE/MODEL-VERSION A-02/16 APR 77
2  C
3  C PURPOSE--
4  C
5  C GIVEN A MATRIX OF COEFFICIENTS 'A' FROM WHICH THE REDUNDANT
6  C COLUMNS AND ROWS HAVE BEEN REMOVED TO FACILITATE ANALYSIS,
7  C PROGRAM USES 'PLODITE' METHOD TO PREPARE A LARGER MATRIX
8  C 'B' CONTAINING THE LEFT-OUT COLUMNS AND ROWS.
9  C
10 C DESCRIPTION--
11 C
12 C PROGRAM CONSTITUTES A GENERAL PLODITE ALGORITHM, IN THAT IT
13 C IS DESIGNED TO ACCEPT AN 'NDIMA' * 'NDIMA' SQUARE MATRIX 'A' OF
14 C ANY SIZE IN WHICH ANY NUMBER OF VARIABLES MAY APPEAR AND IN
15 C WHICH THE LEFT-OUT DUMMY VARIABLES MAY OCCUPY ANY POSITION WITH
16 C IN THE VARIABLE BLOCK ASSIGNED TO THEIR CATEGORY, WITHIN EACH
17 C CATEGORY, HOWEVER, ONE AND ONLY ONE VARIABLE MUST HAVE BEEN
18 C DELETED, EACH CATEGORY MUST HAVE ONE LEFT-OUT DUMMY.
19 C
20 C THE USER SPECIFIES THE NUMBER OF VARIABLE CATEGORIES (E.G.,
21 C CEILING, VISIBILITY, WIND) AND THE NUMBER OF DUMMY VARIABLES
22 C USED TO DESCRIBE EACH SUCH CATEGORY. THE NUMBER OF DUMMIES
23 C MAY VARY FROM ONE CATEGORY TO ANOTHER, THE USER THEN SPECI-
24 C FIES, FOR EACH CATEGORY, WHICH OF THE DUMMIES HAS BEEN LEFT
25 C OUT (E.G., FOR THIRD CATEGORY, DUMMY 4 IS DELETED, FOR
26 C FOURTH CATEGORY, DUMMY 2 IS LEFT OUT, ETC.),
27 C
28 C IN ADDITION TO PARAMETERS SPECIFIED ABOVE, PROGRAM INPUT MUST
29 C INCLUDE A VECTOR EQUAL TO FIRST ROW OF CROUT SUM-OF-SQUARES-
30 C AND-CROSS-PRODUCTS (SSCP) MATRIX, SUPPLIED ON FILE 'IFILE1,'
31 C THIS VECTOR CONTAINS THE NUMBER OF OBSERVATIONS IN ITS
32 C FIRST ELEMENT, SUM OF 'Z1' IN SECOND, SUM OF 'Z2' IN THIRD,
33 C ETC. THE PROGRAM DIVIDES TO GET THE MEANS. THE MATRIX 'A' IS
34 C SUPPLIED ON FILE 'IFILE2' IN TRANSPOSED FORM. IT IS RE-
35 C TRANSPOSED WHILE BEING READ IN.
36 C
37 C PROGRAM OUTPUT IS THE EXPANDED MATRIX 'B' IN ORIGINAL FORM,
38 C NOT TRANSPOSED TO MATCH INPUT. THE 'B' MATRIX IS ALWAYS
39 C PRINTED. IN ADDITION, THE USER MAY SPECIFY THAT THE 'B'
40 C MATRIX BE OUTPUT TO CARDS OR TAPE, AS PROVIDED VIA JCL.
41 C
42 C INTER-COMPUTER COMPATIBILITY IS FACILITATED BY USE OF
43 C VARIABLE FILE CODES AND NON-VENDOR UNIQUE FORTRAN IV.
44 C
45 C INPUT/OUTPUT REQUIREMENTS--
46 C
47 C FILE CONTENTS MEDIA FC NORM VAL
48 C
49 C SYSIN PROBL NAME, PARMS,
50 C VARIABLE STRUC-
51 C TURE SYS DSK IREAD 5
52 C SYSOUT PRINTED OUTPUT SYS DSK IPRINT 6
53 C DATA INPUT 1 SSCP VECTOR TP/DK/CD IFILE1 10
54 C DATA INPUT 2 A MATRIX TP/DK/CD IFILE2 11
55 C ALT OUTPUT B MATRIX TP/DK/CD IQUT 20
56 C
57 C STRUCTURE OF INPUT DATA FILE IREAD--
58 C
59 C CARD COLUMNS FIELD FORMAT DESCRIPTION
60 C
61 C 1 1-72 IPROB(J) 10A4 PROBLEM TITLE, NUMBER, DATE

```

62	C	2	4	IFLGA	11	REQUEST PRINT MTRX A ON IPRINT
63	C		8	IFLGB	11	REQUEST WRITE MTRX B ON IOUT
64	C		12	IFLGC	11	REQUEST PRINT SSCP ON IPRINT
65	C	3	1- 4	NCATEG	14	NUMBER DUMMY VARIABLE CATE-
66	C					GORIES
67	C	4	1- 4	NZ(NC)	14	NUMBER DUMMY VARIABLES IN EACH
68	C		5- 8	NZ(NC)	14	VARIABLE CATEGORY
69	C		9-12	NZ(NC)	14	
70	C		ETC	ETC		
71	C	5	1- 4	NOUT(NC)	14	LOCAL NUMBER INDICATING INDEX
72	C		5- 8	NOUT(NC)	14	OF THE LEFT-OUT VARIABLE IN
73	C		9-12	NOUT(NC)	14	EACH CATEGORY
74	C		ETC	ETC		

STRUCTURE OF INPUT DATA FILE IFILE1--

(TO BE SUPPLIED)

STRUCTURE OF INPUT DATA FILE IFILE2--

(TO BE SUPPLIED)

```

*****
*
* WARNING ... IFILE2 MUST BE IN THE FORM OF THE TRANS-
* POSE OF MATRIX A, WITH 'ZERO-COLUMN' MISSING. THE
* 'ZERO-COLUMN' OF THE TRANSPOSE IS, OF COURSE, THE
* 'ZERO-ROW' OF THE ACTUAL A-MATRIX?
*
*****

```

MODEL-VERSION HISTORY--

MODEL-VERSION A-01, 24 MARCH 1977-- ORIGINAL MODEL-VERSION.

MODEL-VERSION A-02, 16 APRIL 1977-- REVISED METHOD OF INSERT-  
ING LEFT-OUT ROWS OF MATRIX 'B.' EARLIER MODEL-VERSION  
PLACED ROWS LAST IN BLOCK OF ROWS CORRESPONDING TO A VARIABLE  
CATEGORY. NEW VERSION INSERTS THE LEFT-OUT ROW AT ITS  
APPROPRIATE POSITION IN ACCORDANCE WITH NUMBER OF LEFT-OUT  
VARIABLE 'NOUT.' REVISED VERSION ALSO DISPLAYS DUMMY VARIABLE  
COUNT NZ(NC).

PROGRAMMER--

CAPT ROGER C. WHITON

DIMENSION IPROB(18), NZ(3), NOUT(3), A(7,7), B(10,10),  
SSCP(7)

FILE CODES. IREAD IS SYSIN. IPRINT IS PRINTER-DIRECTED  
SYSOUT. 'IFILE1' CONTAINS FIRST ROW OF CROUT SSCP MATRIX ('ZZ'  
IN OTHER PROGRAMS, 'SSCP' IN THIS ONE). 'IFILE2' CONTAINS THE  
MATRIX 'A.' 'IOUT' IS OPTIONAL PUNCH OR TAPE OUTPUT.

```

IREAD   = 5
IPRINT  = 6
IFILE1  = 10
IFILE2  = 11
IOUT    = 20

```

HEADERS. 'IPROB' IS PROBLEM TITLE FOR DISPLAY AS BANNER  
ON OUTPUT.

```

WRITE (IPRINT,6000)
6000 FORMAT (1H1/1X, 'SAINT LOUIS UNIVERSITY',1X, 'DEPARTMENT OF ',
* 'EARTH AND ATMOSPHERIC SCIENCES'///1X, '--PLOGITE MATRIX TREAT',

```



```

130      * 'MENT--'//)
131      READ (IREAD,5000) (IPROB(JX), JX = 1,18)
132      5000 FORMAT (18A4)
133      WRITE (IPRINT,6010) (IPROB(JX), JX = 1,18)
134      6010 FORMAT (1H0, 'PROBLEM-- ', T50, 18A4)
135      C
136      C READ FLAGS FOR DETERMINING OUTPUT MEDIA. 'B' MATRIX IS ALWAYS
137      C PRINTED, BUT IF 'IFLGB' IS ON, ADDITIONAL OUTPUT TO FILE
138      C 'IOUT' IS MADE. IF 'IFLGA' IS ON, MATRIX 'A' IS ALSO OUTPUT.
139      C IF 'IFLGC' IS ON, CROUT SSCP VECTOR IS DISPLAYED.
140      C
141      READ (IREAD,5010) IFLGA, IFLGB, IFLGC
142      C
143      C MATRIX OF COEFFICIENTS A(I,J) IN LEFT-OUT VARIABLE FORM IS CON-
144      C SIDERED PARTITIONED INTO A 20-ROW (1 IN FIRST POSITION FOLLOWED
145      C BY ZEROES), AN A0-COLUMN (LEFTMOST COLUMN), AND THEN
146      C NCLASS * NCLASS PARTITIONS OF COEFFICIENTS. 'NCATEG' IS THE
147      C NUMBER OF DUMMY VARIABLE CATEGORIES IN THE REGRESSION SCHEME.
148      C FOR EXAMPLE, A SCHEME IN CEILING, VISIBILITY AND WIND (5
149      C CLASSES OF CEILING, 6 CLASSES OF VISIBILITY, 9 CLASSES OF WIND)
150      C WOULD HAVE NCATEG = 3. 'NC' IS INDEX OF CATEGORIES, NC = 1,
151      C NCATEG. FOR EACH SUCH CATEGORY, NZ(NC) GIVES NUMBER OF DUMMY
152      C VARIABLE CLASSES (INCLUDING LEFT-OUTS) INTO WHICH THE NC-TH
153      C CATEGORY IS SUBDIVIDED. IN THE EXAMPLE ABOVE, NZ(1) = 5,
154      C NZ(2) = 6, NC(3) = 9. READ 'NCATEG' AND 'NZ.'
155      C
156      READ (IREAD,5010) NCATEG
157      5010 FORMAT (18I4)
158      READ (IREAD,5010) (NZ(NC), NC = 1,NCATEG)
159      C
160      C READ WHICH VARIABLES NOUT(NC) ARE LEFT OUT FROM A(I,J). 'NOUT'
161      C IS THE NOUT-TH VARIABLE IN THE NC-TH CATEGORY. READ 'NOUT,'
162      C CHECKING TO SEE IT DOES NOT EXCEED 'NZ.'
163      C
164      READ (IREAD,5010) (NOUT(NC), NC = 1,NCATEG)
165      DO 100 NC = 1,NCATEG
166      IF (NOUT(NC) .LE. NZ(NC)) GO TO 100
167      WRITE (IPRINT,6020) NC, NOUT(NC), NZ(NC)
168      6020 FORMAT (1H0, 'ERROR... ', I4, 'TH LEFT-OUT VARIABLE NUMBER WAS ',
169      * I4/1X, 'EXCEEDS NUMBER OF DUMMY VARIABLES (', I4, ') IN ',
170      * 'CATEGORY'/1X, 'PROGRAM STOPS')
171      STOP
172      100 CONTINUE
173      C
174      C THERE MUST BE AT LEAST ONE LEFT-OUT VARIABLE IN EACH CATEGORY.
175      C
176      DO 200 NC = 1,NCATEG
177      IF (NOUT(NC) .GT. 0) GO TO 200
178      WRITE (IPRINT,6030) NC, NOUT(NC)
179      6030 FORMAT (1H0, 'ERROR... ', I4, 'TH LEFT-OUT VARIABLE NUMBER WAS ',
180      * I4/1X, 'SHOULD BE GREATER THAN ZERO'/1X, 'PROGRAM STOPS')
181      STOP
182      200 CONTINUE
183      C
184      C CALCULATE SIZE OF LEFT-OUT MATRIX 'A' AND RESTORED MATRIX 'B'
185      C AND DISPLAY RESULTS. 'NVRBLZ' IS NUMBER OF DUMMIES, INCLUDING
186      C Z0. 'NVRBOT' IS NUMBER OF LEFT-OUT VARIABLES. 'NDIMA' IS
187      C DIMENSIONALITY OF MATRIX 'A,' AND 'NDIMB' IS DIMENSIONALITY OF
188      C MATRIX 'B.'
189      C
190      NVRBLZ = 1
191      DO 300 NC = 1,NCATEG
192      NVRBLZ = NVRBLZ + NZ(NC)
193      300 CONTINUE
194      NVRBOT = NCATEG
195      NDIMB = NVRBLZ
196      NDIMA = NCIMB - NVRBOT
197      WRITE (IPRINT,6040) NCATEG, NVRBLZ-1, NVRBOT, NDIMA, NDIMA,
198      * NDIMB, NDIMB

```

```

199 6040 FORMAT (1H , 'NUMBER DUMMY VARIABLE CATEGORIES-- ', T48, 14/1X,
200   * 'NUMBER DUMMY VARIABLES EXCLUDING ZERO-- ', T48, 14/1X,
201   * 'NUMBER OF LEFT-OUT DUMMY VARIABLES-- ', T48, 14/1X,
202   * 'DIMENSIONALITY OF LEFT-OUT MATRIX A-- ', T48, 14, ' * ', 14/1X,
203   * 'DIMENSIONALITY OF PLODITE OUTPUT MATRIX B-- ', T48, 14, ' * ',
204   * 14///)
205 C
206 C   DISPLAY NUMBER OF DUMMY VARIABLES IN EACH CATEGORY.
207 C
208   WRITE (IPRINT,6043)
209 6043 FORMAT (1H0, 'DUMMY VARIABLES NZ(NC) IN VARIABLE CATEGORY NC--')//
210   * 1X, 10(4X, 'NC', 2X, ' NZ ')//)
211   WRITE (IPRINT,6046) (NC, NZ(NC), NC = 1, NCATEG)
212 C
213 C   DISPLAY LEFT-OUT DUMMY VARIABLE NUMBERS.
214 C
215   WRITE (IPRINT,6045)
216 6045 FORMAT (1H0, 'LEFT-OUT DUMMY VARIABLES NOUT(NC) IN VARIABLE ',
217   * 'CATEGORY NC-- '///1X, 10(4X, 'NC', 2X, 'NOUT')//)
218   WRITE (IPRINT,6046) (NC, NOUT(NC), NC = 1, NCATEG)
219 6046 FORMAT (4X, 13, 2X, 13, 4X, 13, 2X, 13, 4X, 13, 2X, 13, 4X, 13,
220   * 2X, 13, 4X, 13, 2X, 13, 4X, 13, 2X, 13, 4X, 13, 2X, 13, 4X,
221   * 13, 2X, 13, 4X, 13, 2X, 13, 4X, 13, 2X, 13)
222 C
223 C   READ FIRST ROW OF CROUT SSCP MATRIX. FIRST ELEMENT IS NUMBER
224 C   OF CASES IN SAMPLE. NEXT IS SUM OF 'Z1', NEXT SUM OF 'Z2', ETC.
225 C   DIVIDE BY NUMBER OF CASES TO GET MEANS. NOTE THAT THIS
226 C   CROUT IS IN LEFT-OUT VARIABLE FORM.
227 C
228   READ (IFILE1,5015) (SSCP(JX), JX = 1,NDIMA)
229 5015 FORMAT (6F12.0)
230   DO 400 JX = 2,NDIMA
231     SSCP(JX) = SSCP(JX) / SSCP(1)
232 400 CONTINUE
233 C
234 C   DISPLAY CROUT VECTOR IF 'IFLGC' IS ON.
235 C
236   IF (IFLGC .LE. 0) GO TO 450
237   WRITE (IPRINT,6050)
238 6050 FORMAT (1H0, 'FIRST ROW OF SUM OF SQUARES AND CROSS PRODUCTS ',
239   * 'MATRIX (N, MEANS)-- '///)
240   WRITE (IPRINT,6060)
241 6060 FORMAT (1H , 4X, 'NV', 3X, 'SIGMA V', 8X, 'NV', 3X,
242   * 'SIGMA V', 7X, 'NV', 3X, 'SIGMA V', 8X, 'NV', 3X, 'SIGMA V',
243   * 8X, 'NV', 3X, 'SIGMA V'//)
244   WRITE (IPRINT,6070) (JX, SSCP(JX), JX = 1,NDIMA)
245 6070 FORMAT (1H , 2X, 14, 1X, E12.5, 4X, 13, 1X, E12.5, 3X, 13, 1X,
246   * E12.5, 4X, 13, 1X, E12.5, 4X, 13, 1X, E12.5)
247 C
248 C   READ MATRIX A(I,J). IT HAS BEEN PREPARED IN THE FORM OF TRANS-
249 C   POSE, SO READ IT IN TRANSPOSED FORM SUCH THAT WHEN A(I,J)
250 C   RESIDES IN CORE IT WILL BE IN ORIGINAL FORM. ROWS 'I' WILL
251 C   CORRESPOND TO THE REGRESSION EQUATIONS, AND COLUMNS 'J' WILL
252 C   CORRESPOND TO VARIABLES IN THE EQUATIONS. NOTE THAT THE 'A'
253 C   MATRIX SUPPLIED ON 'IFILE2' IS MISSING THE 'ZERO-ROW.' THUS
254 C   THE READING IS INTO AREAS OF THE MATRIX BEYOND THE ZERO-ROW.
255 C   THE ZERO ROW IS LATER SUPPLIED.
256 C
257 450 READ (IFILE2,5020) ((A(I,J), I = 2,NDIMA), J = 1,NDIMA)
258 5020 FORMAT (6E13.5, 2X)
259 C
260 C   SUPPLY ZERO-ROW.
261 C
262   DO 525 J = 1,NDIMA
263     A(1,J) = 0.0
264 525 CONTINUE
265     A(1,1) = 1.0
266 C
267 C   ELEMENT A(1,1) SHOULD BE UNITY AND TAKES ROLE OF A(0,0). REST

```

```

268 C      OF ROW 1 SHOULD BE ZEROES. COPY THIS INTO 'B,'
269 C
270 B(1,1) = A(1,1)
271 WRITE (IPRINT,6080) B(1,1)
272 6080 FORMAT (1H0, 'FIRST OR ''ZERO''-ROW-- '//5X, 'B(1,1) IS ',
273 * E12,5/5X, 'ALL OTHER B(1,J) SET TO ZERO')
274 DO 600 J = 2,NDIMB
275 B(1,J) = 0.0
276 600 CONTINUE
277 C
278 C      DISPLAY MATRIX 'A' IF 'IFLGA' IS ON.
279 C
280 IF (IFLGA .LE. 0) GO TO 650
281 WRITE (IPRINT,6120)
282 6120 FORMAT (1H1, 'ORIGINAL MATRIX ''A'' IN LEFT-OUT VARIABLE ',
283 * 'FORM-- '//)
284 WRITE (IPRINT,6130)
285 6130 FORMAT (1H, 'I', 'T9, 'J', 'T15, 'A(J)', 6(7X, 'J', 5X,
286 * 'A(J)'))
287 DO 625 I = 1,NDIMA
288 WRITE (IPRINT,6100) I
289 WRITE (IPRINT,6105) (J, A(I,J), J = 1,NDIMA)
290 625 CONTINUE
291 C
292 C      I-LOOP GOES DOWN THROUGH 'NCATEG' BLOCKS OF ROWS IN BOTH 'A'
293 C      AND 'B.'
294 C
295 650 DO 2500 ICATEG = 1,NCATEG
296 C
297 C      'IPOINT' POINTS TO FIRST ROW OF 'B' IN PARTITION BEING TREATED.
298 C      'IPTSS' POINTS TO INDEX OF 'SSCP' AND 'A' ARRAYS CORRESPONDING TO
299 C      PROPER ELEMENT OF 'B.'
300 C
301 660 IPOINT = 2
302 IPTSS = 2
303 IF (ICATEG .EQ. 1) GO TO 705
304 DO 700 IX = 1,ICATEG-1
305 IPOINT = IPOINT + NZ(IX)
306 670 IPTSS = IPOINT - IX
307 700 CONTINUE
308 C
309 C      'IVOUT' POINTS TO LEFT-OUT ROW OF B-MATRIX APPLICABLE IN PRESENT
310 C      BLOCK OF I-ROWS. IT MUST BE COMPUTED ONCE FOR EACH I-ROW
311 C      BLOCK CORRESPONDING TO VARIABLE CATEGORY.
312 C
313 705 IVOUT = 1
314 IF (ICATEG .EQ. 1) GO TO 715
315 DO 710 IX = 1,ICATEG-1
316 IVOUT = IVOUT + NZ(IX)
317 710 CONTINUE
318 715 IVOUT = IVOUT + NOUT(ICATEG)
319 C
320 C      ADVANCE THROUGH AN I-BLOCK. 'I' IS ABSOLUTE INDEX IN 'B'
321 C      ARRAY.
322 C
323 725 DO 2000 I = IPOINT,IPOINT+NZ(ICATEG)-1
324 C
325 C      SKIP THE LEFT-OUT ROW 'IVOUT,'
326 C
327 IF (I .EQ. IVOUT) GO TO 2000
328 C
329 C      FIND THE APPROPRIATE ROW OF 'A' FROM WHICH TO TAKE THE COEFFI-
330 C      CIENTS USED IN PREPARING LEFT-OUT 'B': THIS NEED BE DONE
331 C      ONLY ONCE FOR EACH 'I,'
332 C
333 740 IA = 1
334 IF (ICATEG .EQ. 1) GO TO 775
335 DO 770 IX = 1,ICATEG-1
336 IA = IA + (NZ(IX) - 1)

```



```

337 770 CONTINUE
338 DO 780 IX = IPOINT, I
339 IF (IX .EQ. IVOUT) GO TO 780
340 IA = IA + 1
341 780 CONTINUE
342 C
343 C J-LOOP GOES ACROSS THROUGH 'NCATEG' BLOCKS OF COLUMNS. FIRST
344 C COLUMN IS ALWAYS OMITTED BECAUSE IT IS TREATED SEPARATELY
345 C BELOW.
346 C
347 C ZERO THE B-ACCUMULATOR ('BACCUM') USED TO SET THE 'B0' COLUMN.
348 C
349 C BACCUM = 0.0
350 C
351 C DO 1500 JCATEG = 1, NCATEG
352 C
353 C 'JPOINT' POINTS TO FIRST COLUMN OF 'B' IN PARTITION BEING
354 C TREATED.
355 C
356 790 JPOINT = 2
357 IF (JCATEG .EQ. 1) GO TO 825
358 DO 800 JX = 1, JCATEG-1
359 JPOINT = JPOINT + NZ(JX)
360 800 CONTINUE
361 C
362 C 'JENDB' IS END OF J-SCAN IN B-ARRAY.
363 C
364 825 JENDB = JPOINT + NZ(JCATEG) - 1
365 C
366 C FIRST SUPPLY THE 'B' CORRESPONDING TO THE LEFT-OUT 'A.' THE
367 C NOUT(JCATEG)-TH A-VALUE IS LEFT OUT. COMPUTE ABSOLUTE INDEX
368 C 'JOUT' OF LEFT-OUT 'B.'
369 C
370 C JOUT = JPOINT + NOUT(JCATEG) - 1
371 C
372 C COMPUTE INDEX LIMITS IN A-ARRAY FOR A-VALUES TO BE USED IN COM-
373 C PUTING THE LEFT-OUT 'B.' THE TRICK IS THAT ALL OF THE A-
374 C VALUES WILL ALWAYS BE USED IN THIS STEP, ONE NEED ONLY MULTIPLY
375 C THEM BY APPROPRIATE SSCP ELEMENTS, ADD, AND CHANGE SIGN,
376 C
377 840 JBGNA = 2
378 JENDA = JBGNA + (NZ(1) - 2)
379 IF (JCATEG .EQ. 1) GO TO 900
380 DO 850 JA = 1, JCATEG-1
381 JINC = (NZ(JA) - 1)
382 JBGNA = JBGNA + JINC
383 JENDA = JBGNA + NZ(JA+1) - 2
384 850 CONTINUE
385 C
386 C SELECTING A'S FROM ROW 'IA' AND BETWEEN 'JBGNA' AND 'JENDA' AND
387 C Z'S FROM CORRESPONDINGLY J-INDEXED SSCP-VECTOR, PERFORM MULTI-
388 C PPLICATION TO PREPARE THE LEFT-OUT 'B.'
389 C
390 900 B(1, JOUT) = 0.0
391 DO 950 JA = JBGNA, JENDA
392 B(1, JOUT) = B(1, JOUT) + ( A(IA, JA) * SSCP(JA) )
393 950 CONTINUE
394 B(1, JOUT) = -B(1, JOUT)
395 C
396 C SAVE THIS LEFT-OUT 'B' AS CORRECTION TERM TO BE APPLIED TO THE
397 C OTHER A'S IN ITH ROW OF B-ARRAY.
398 C
399 960 Q = B(1, JOUT)
400 C
401 C ACCUMULATE ALL LEFT-OUT B'S FOR THE ROW 'I' BEING TREATED. THE
402 C ACCUMULATED TOTAL WILL LATER BE USED IN PREPARING ELEMENTS
403 C B(I, 1).
404 C

```

```

405      965 BACCUM = BACCUM + 0
406      C
407      C NOW THE TASK IS TO FILL THE REMAINING B'S IN ROW 'I' FOR
408      C J = JPOINT,JENDB FOR J NOT EQUAL TO 'JOUT' (ALREADY COMPUTED,
409      C SKIP),
410      C
411      C FIND JA-INDEX OF ARRAY 'A' CORRESPONDING TO THE 'JPOINT' ELEMENT
412      C OF ARRAY 'B.' THEN EXTRACT NZ(JCATEG)-1 A-VALUES FROM ARRAY,
413      C CORRECT THEM WITH 0, AND STORE IN ASCENDING LOCATIONS OF 'B,'
414      C SKIPPING THE 'JOUT' ELEMENT ALREADY TREATED.
415      C
416      970 JA = 1
417      IF (JCATEG .EQ. 1) GO TO 1000
418      DO 980 JX = 1,JCATEG-1
419      JA = JA + (NZ(JX) - 1)
420      980 CONTINUE
421      1000 J = JPOINT
422      1010 JA = JA + 1
423      IF (JA .GT. NDIMA) GO TO 1500
424      TEMPO = A(JA,JA) + 0
425      IF (J .EQ. JOUT) J = J + 1
426      IF (J .GT. JENDB) GO TO 1500
427      B(I,J) = TEMPO
428      J = J + 1
429      IF (J .GT. JENDB) GO TO 1500
430      GO TO 1010
431      C
432      C PROGRAM BRANCHES TO HERE WHEN A J-CATEGORY IS COMPLETE. THIS
433      C IS TERMINATION OF J-LOOP.
434      C
435      1500 CONTINUE
436      C
437      C WHEN ALL J-CATEGORIES FOR A GIVEN I-ROW ARE COMPLETE, THE I-
438      C ROW ADVANCE LOOP TERMINATES HERE. NEW I-ROW IN BLOCK OF I-
439      C ROWS CONSTITUTING PRESENT I-CATEGORY.
440      C
441      C BEFORE GOING TO NEW I-ROW, SUPPLY THE 'B0' TERM FOR THE ROW
442      C BEING TREATED. THIS IS SUM OF LEFT-OUT B'S (BACCUM) SUBTRACTED
443      C FROM THE 'A' CORRESPONDING TO THIS I-ROW. IN OTHER WORDS,
444      C THE B(I,1) TERMS ARE EQUAL TO THE MEANS OF THE ASSOCIATED
445      C PREDICTANDS (NOT PREDICTORS). 'IA' HAS ALREADY BEEN COM-
446      C PUTED AS THE I-INDEX OF 'A' CORRESPONDING TO THE I-INDEX
447      C OF 'B.'
448      C
449      1900 B(I,1) = A(IA,1) - BACCUM
450      2000 CONTINUE
451      C
452      C WHEN ALL I-ROWS IN PRESENT I-CATEGORY ARE COMPLETE, I-LOOP TER-
453      C MINATES HERE FOR A NEW I-CATEGORY (NEW BLOCK OF I-ROWS).
454      C FIRST WE MUST TREAT THE IVOUT-TH I-ROW AS NEGATIVE OF SUM OF
455      C THOSE BEFORE IT. 'B0' COLUMN ADDS TO ONE. OTHERS TO ZERO.
456      C
457      2020 IX = IVOUT
458      DO 2050 J = 1,NDIMB
459      B(IX,J) = 0.0
460      2050 CONTINUE
461      DO 2100 J = 1,NDIMB
462      DO 2090 I = JPOINT,JPOINT+NZ(JCATEG)-1
463      IF (I .EQ. IVOUT) GO TO 2090
464      2070 B(IX,J) = B(IX,J) + B(I,J)
465      2090 CONTINUE
466      2095 B(IX,J) = -B(IX,J)
467      2100 CONTINUE
468      2110 B(IX,1) = 1.0 + B(IX,1)
469      2500 CONTINUE
470      C
471      C WHEN PROGRAM CONTROL ARRIVES AT THIS POINT, THE B-MATRIX IS COM-
472      C PLETE AND READY FOR OUTPUT. ALWAYS PRINT THE MATRIX. OUTPUT
473      C TO FILE 'IOUT' IS OPTIONAL, DEPENDING ON OUTPUT FLAG 'IFLGB.'

```

```

474 C      IF WRITTEN, 'IOUT' CAN BE ADDRESSED TO TAPE, DISK OR CARDS, AS
475 C      PROVIDED IN JCL.
476 C
477      WRITE (IPRINT,6090)
478 6090 FORMAT (1H1, 'PLODITE OUTPUT MATRIX ..B...-- .//')
479      WRITE (IPRINT,6095)
480 6095 FORMAT (1H, 'T4, 'I', T9, 'J', T15, 'B(J)', 6(7X, 'J', 5X,
481      * 'B(J)'))
482      DO 3000 I = 1,NDIMB
483      WRITE (IPRINT,6100) I
484 6100 FORMAT (1H0, T2, I3)
485      WRITE (IPRINT,6105) (J, B(I,J), J = 1,NDIMB)
486 6105 FORMAT (T5, 2X, I3, 1X, E11.4, 2X, I3, 1X, E11.4, 2X, I3, 1X,
487      * E11.4, 2X, I3, 1X, E11.4, 2X, I3, 1X, E11.4, 2X, I3, 1X, E11.4,
488      * 2X, I3, 1X, E11.4)
489 3000 CONTINUE
490      IF (IFLGB,LE. 0) GO TO 4000
491      DO 3100 I = 1,NDIMB
492      WRITE (IOUT,6110) (B(I,J), J = 1,NDIMB)
493 6110 FORMAT (5E15.8)
494 3100 CONTINUE
495      REWIND IOUT
496 C
497 C      TERMINATION.
498 C
499 4000 STOP
500      END

```



# Appendix B

## CROUT PROGRAM

C THIS PROGRAM PRODUCES REGRESSION COEFFICIENTS USING THE CROUT  
C FORWARD-BACKWARD SOLUTION METHOD

```

0001      DIMENSION SSR( 20)
0002      REAL*8 A( 20, 20),B( 20, 20),Y( 20, 20)

C
C READ N, THE ORDER OF THE SQUARE MATRIX A. NM, THE NUMBER OF
C PREDICTANDS APPEARING IN THE SSCP PREDICTOR-PREDICTAND MATRIX.
C
0003      READ(5,900) N,NM
0004      900  FORMAT(2I3)
0005      NP=N+1
0006      NL=N-1

C
C READ IN THE SSCP PREDICTOR MATRIX
C
0007      READ(5,901) ((A(I,J),J=1,N),I=1,N)
C
C READ THE SSCP PREDICTOR - PREDICTAND MATRIX.
C
0008      READ(5,901) ((Y(I,J),J=1,N),I=1,NM)
0009      901  FORMAT( 7F5.0)

C
C BEGIN CALCULATING THE CROUT AUXILIARY MATRIX BY GETTING THE FIRST ROW
C
0010      DO 10 I=2,N
0011      10  A(I,1)=A(I,1)/A(1,1)

C
C COMPLETE THE CROUT AUXILIARY MATRIX
C
0012      DO 20 J=2,N
0013      DO 30 I=J,N
0014      JS=J-1
0015      DO 40 L=1,JS
0016      A(I,J)=A(I,J)-A(I,L)*A(L,J)
0017      40  CONTINUE
0018      IF(J.EQ.1) GO TO 30
0019      A(J,1)=A(I,J)/A(J,J)
0020      30  CONTINUE
0021      20  CONTINUE

C
C K REPRESENTS THE VARIABLE FOR WHICH YOU ARE DERIEIVING COEFFICIENTS
C
0022      DO 500 K=1,NM

C
C AUGMENT THE CROUT AUXILIARY MATRIX BY MAKING A LAST ROW WITH THE
C K TH ROW OF THE SSCP PREDICTOR-PREDICTAND MATRIX.
0023      DO 50 M=1,N
0024      50  A(NP,M)=Y(K,M)

C
C PLACE THE 1 ST ELEMENT OF THE K TH ROW INTO THE LAST DIAGONAL ELEMENT
C OF THE AUGMENTED AUXILIARY
C
0025      A(NP,NP)=Y(K,1)

C
C START TO CALCULATE THE ADDITIONAL ROW AND COLUMN
C
0026      A(1,NP)=A(NP,1)/A(1,1)

C
C COMPLETE PROCESSING THE ADDITIONAL ROW AND COLUMN AS IN 20 LOOP ABOVE
C
0027      DO 200 J=2,NP
0028      JS=J-1
0029      DO 400 L=1,JS
0030      400  A(NP,J)=A(NP,J)-A(NP,L)*A(L,J)
0031      IF(J.EQ.NP) GO TO 200
0032      A(J,NP)=A(NP,J)/A(J,J)
0033      200  CONTINUE

```

```

C
C BEGIN CALCULATING THE COEFFICIENTS FOR THE K TH VARIABLE BY GETTING THE
C LAST COEFFICIENT
C
0034      B(K,N)=A(N,NP)
0035      DO 450 J=1,NL
0036      JJ=N-J
0037      B(K,JJ)=A(JJ,NP)
0038      DO 300 L=1,J
0039      LL=NP-L
0040      300  B(K,JJ)=B(K,JJ)-A(JJ,LL)*B(K,LL)
0041      450  CONTINUE
C
C CALCULATE THE RESIDUAL SUM OF SQUARES
C
0042      SSR(K)=A(NP,NP)/A(1,1)
0043      500  CONTINUE
C
C OUTPUT THE COEFFICIENTS
C
0044      DO 600 K=1,NM
0045      WRITE(6,902)K,SSR(K)
0046      902  FORMAT(1H0,I6,E14.6)
0047      WRITE(6,903) (J,B(K,J),J=1,N)
0048      903  FORMAT(6(15,E15.6))
0049      600  CONTINUE
0050      STOP

```

# Appendix C

## REGRESSION ESTIMATION OF EVENT PROBABILITIES

(REEP)

### C DESCRIPTION

C PERFORMS REEP ANALYSIS( SEE CHAPTER 5 ) GIVEN THE CROSSPRODUCT MATRIX AMONG A SET OF PREDICTORS(ZZ) AND THE CROSSPRODUCT MATRIX BETWEEN A SET OF PREDICTANDS AND PREDICTORS(YZ).  
C SCREENING IS PERFORMED AMONG THE PREDICTORS TO MAXIMIZE THE PREDICTABILITY FOR A GROUP OF PREDICTANDS. ALL OF THE PREDICTANDS IN A GROUP ARE DUMMY VARIABLES, USUALLY MUTUALLY EXCLUSIVE AND EXHAUSTIVE.  
C NOTE THAT THE ORDINARY F TEST IS NOT APPROPRIATE WHEN SELECTING PREDICTORS BY A SCREENING PROCEDURE. IT IS RECOMMENDED THAT THE PROBABILITY LEVEL OF THE TEST BE MADE A FUNCTION OF THE NUMBER OF POSSIBLE PREDICTORS(IP), SPECIFICALLY,  $(1 - 1/(20*IP))$ . THUS AN ORDINARY 95% LEVEL WITHOUT SCREENING WOULD BE  $(1-1/20)$ . SINCE MOST STATISTICAL TABLES SUCH AS HALD'S DO NOT COVER THE SITUATION OF A LARGE NUMBER OF PREDICTORS, IT IS SUGGESTED THAT THE INVERTED PAULSON APPROXIMATION BE USED TO ARRIVE AT THE CRITICAL F VALUE. A FACTOR W IS USED TO APPROPRIATELY REDUCE THE NUMBER OF SAMPLE CASES BECAUSE OF SERIAL CORRELATION. FOR EXAMPLE, IF ONLY EVERY EIGHTH OBSERVATION CAN BE CONSIDERED INDEPENDENT, SET W EQUAL TO 8.

$$C \quad \text{INVERTED PAULSON F APPROXIMATION: } F_{\epsilon}(n_1, n_2) = \left[ \frac{ab + \sqrt{a^2b^2 - [a^2(1-a)K^2][b^2(1-b)K^2]}}{[b^2(1-b)K^2]} \right]^3$$

C  $a = 1 - \frac{2}{9n_1}$  and  $b = 1 - \frac{2}{9n_2}$

C WHERE K IS THE NUMBER OF STANDARD DEVIATIONS THE ACCUMULATED PROBABILITY  $\epsilon$  IS FROM THE MEAN OF THE NORMAL DISTRIBUTION.

### C INPUT

CARDS	COL	FMT	NAME	DESCRIPTION
1	1-3	I3	IP	NUMBER OF PREDICTORS. ZZ IS AN IP X IP MATRIX
	4-6	I3	MM	NUMBER OF PREDICTANDS. YZ IS AN MM X IP MATRIX
	7-9	I3	IG	NUMBER OF PREDICTAND GROUPS. ( E.G. CEILING, VISIBILITY, PRESSURE, ETC)
	10-12	I3	IGG	MAXIMUM NUMBER OF PREDICTAND DUMMY CATEGORIES IN ANY OF THE IG PREDICTAND GROUPS.
2	1-3	I3	N(1,IG)	DEFINES THE START POINT FOR EACH OF THE IG PREDICTAND GROUPS.
	4-5	I3	N(2,IG)	DEFINES THE END POINT FOR EACH OF THE IG PREDICTAND GROUPS.
				FOR EXAMPLE, SUPPOSE CEILING IS GROUP 2 AND ITS DUMMY PREDICTANDS ARE FROM 20 TO 26 IN THE YZ MATRIX. THEN N(1,2)=20 AND N(2,2)= 26. THERE WILL BE IG CARDS OF THIS TYPE.
IG+2	1-10	F5.1	FCRIT	CRITICAL F VALUE.
	6-10	F5.1	W	FACTOR THAT IS DIVIDED INTO THE NUMBER OF

C OBSERVATIONS TO GET THE NUMBER OF INDEPENDENT OBS.

C ZZ AND YZ MUST ALSO BE INPUT AND THIS IS DONE VIA SUBROUTINE SCRNIIP.  
C IN THE SCRNIIP LISTING PROVIDED, ZZ AND YZ ARE READ FROM TAPE. ZZ IS A 131X 131 MATRIX AND YZ IS A 131 X 130 MATRIX WHICH MUST BE TRANSPOSED.

C PROGRAMMER - DR ROBERT G. MILLER AND CAPT MICHAEL KELLY



```

        DIMENSION ZZ(131,131),YZ(131,131),B(25,132),Z(132,132),Y(25,132),
        *N(2,22),IS(132),TEMP(132),TRC(25),TRP(25),TRB(25),BEST(132),
        *SSR(25)
        DATA TRC,TRP,TRB/75*0./
        DATA BEST/132*0./
        READ(5,900) IP,MM,IG,IGG
900    FORMAT(4I3)
        READ(5,901) ((N(I,J),I=1,2),J=1,IG)
901    FORMAT(2I3)
        READ(5,902) FCRIT,W
902    FORMAT(2F5.1)
        C SCRNP IS AN INPUT SUBROUTINE. IP AND MM ARE PASSED TO DEFINE THE
        C SIZE OF ZZ AND YZ, THE CROSS PRODUCT MATRICES MENTIONED ABOVE.
        CALL SCRNP(IP,MM,ZZ,YZ)
        ITT=0
        NP=0
        IS(1)=1
        C RESTART POINT FOR NEW PREDICTAND GROUP
1000   IT=0
        C L DEFINES THE PREDICTOR BEING CONSIDERED.
        L=1
        Z(1,1)=ZZ(1,1)
        C NP INDICATES THE PREDICTAND GROUP YOU ARE WORKING WITH
        NP=NP+1
        C KB AND KE DEFINE LIMITS OF THE NP PREDICTAND GROUP
        KB=N(1,NP)
        KE=N(2,NP)
        KSIZE=KE-KB+1
        DC 1100 M=KB,KE
        IT=IT+1
        Y(IT,1)=YZ(M,1)
        C TRANSFORM TO SUM OF SQUARES OF DEVIATIONS FROM THE MEAN.
        TRP(IT)=Y(IT,1)-Y(IT,1)**2/Z(1,1)
1100   CONTINUE
        C RESTART POINT FOR SELECTING NEXT PREDICTOR
2200   IT=1
        X=0.
100    DO 55 J=2,IP
        C TEST TO SEE IF PREDICTOR WAS PREVIOUSLY SELECTED.

        DO 5 I=1,L
        IF(J-IS(I)) 5,55,5
5        CONTINUE
        C LOAD TEMP WITH POSSIBLE PREDICTOR
        DO 10 I=1,L
        K=IS(I)
        TEMP(I)=ZZ(K,J)
10        CONTINUE
        LT=L+1
        C LOAD SUM OF SQUARES OF PREDICTOR INTO TEMP, A WORK SPACE.
        TEMP(LT)=ZZ(J,J)
        C LOAD Y
        DO 15 M=KB,KE
        ITT=ITT+1
        Y(ITT,LT)=YZ(M,J)
        ITI=IS(L)
        Y(ITT,L)=YZ(M,ITI)
15        CONTINUE
        ITT=0
        C TEST TO DETERMINE IF FIRST PREDICTOP
        IF(L-2) 27,28,28

```

```

C AUGMENT TEMP
28   LLT=LT-1
      DO 20 LL=2,LLT
        LLN=LL-1
        DO 25 LN=1,LLN
          TEMP(LL)=TEMP(LL)-TEMP(LN)*Z(LN,LL)
25   CONTINUE
20   CONTINUE
27   DO 22 LN=1,L
      TEMP(LT)=TEMP(LT)-TEMP(LN)**2/Z(LN,LN)
22   CONTINUE
      IT=KSZE
      IF(L-2) 37,38,38
38   DO 30 M=1,IT
      DO 35 LL=L,L
        KI=LL-1
C AUGMENT Y
      DO 31 KK=1,KI
        Y(M,LL)=Y(M,LL)-Y(M,KK)*Z(KK,LL)
31   CONTINUE
35   CONTINUE
30   CONTINUE
37   DO 32 M=1,IT
      DO 34 LL=1,L
        Y(M,LT)=Y(M,LT)-Y(M,LL)*TEMP(LL)/Z(LL,LL)
34   CONTINUE
32   CONTINUE

      DO 40 M=1,IT
C DIAGONAL TEST
      IF(TEMP(LT)-1.E-6) 55,55,43
43   TF=Y(M,LT)**2/TEMP(LT)
      TRC(M)=TRP(M)-TF
C TEST FOR MAX RATIO
      IF(TF/TRC(M)-X) 40,40,45
45   X=TF/TRC(M)
      IJ=J
      DO 47 KK=1,LT
C TRANSFER TEMP INTO BEST
      BEST(KK)=TEMP(KK)
47   CONTINUE
      DO 400 MM=1,IT
        TRB(MM)=TRP(MM)-Y(MM,LT)**2/TEMP(LT)
400  CONTINUE
40   CONTINUE
55   CONTINUE
C DO SIGNIFICANCE TEST
      IF((X*ZZ(L,L)/W-LT)-FCRIT) 2000,2000,2100
2100 ITT=0
      L=L+1
C PREDICTOR IS SIGNIFICANT- INCORPORATE INTO Z.
      DO 550 KK=1,LT
        Z(L,KK)=BEST(KK)
550  CONTINUE
      IIT=LT-1
      DO 420 KK=1,IIT
        Z(KK,L)=BEST(KK)/Z(KK,KK)
420  CONTINUE
      IS(L)=IJ
      DO 450 M=1,IT
        TRP(M)=TRB(M)
450  CONTINUE
C TEST TO SEE IF LAST PREDICTOR.
      IF(L-IP) 2200,2200,2000

```

C PERFORM BACK SOLUTION

2000 M=NP+1

LT=L+1

DO 5000 K=1,KSZE

C TACK ADDITIONAL ROW ON Z

DO 5010 J=1,L

Z(LT,J)=Y(K,J)

5010 CONTINUE

C TACK ADDITIONAL COL ON Z

DO 5020 J=1,L

Z(J,LT)=Y(K,J)/Z(J,J)

5020 CONTINUE

Z(LT,LT)=Y(K,1)

C DETERMINE THE RESIDUAL SUM OF SQUARES

DO 5030 J=1,L

Z(LT,LT)=Z(LT,LT)-Z(LT,J)\*Z(J,LT)

5030 CONTINUE

C BEGIN CALCULATING THE COEFFICIENTS

NS=LT

NL=L-1

B(K,L)=Z(L,NS)

DO 5040 J=1,NL

JJ=L-J

B(K,JJ)=Z(JJ,NS)

DO 5050 I=1,J

LS=NS-I

B(K,JJ)=B(K,JJ)-Z(JJ,LS)\*B(K,LS)

5050 CONTINUE

5040 CONTINUE

C CALCULATE THE RESIDUAL VARIANCE

SSR(K)=Z(NS,NS)/Z(1,1)

5000 CONTINUE

C OUTPUT THE COEFFICIENTS

WRITE(6,904) NP

904 FORMAT('1 THIS IS PREDICTAND GROUP NUMBER ',I3)

DO 6000 K=1,KSZE

WRITE(6,905) K,SSR(K)

905 FORMAT('0 THE RESIDUAL VARIANCE FOR THE ',I3,

\*' PREDICTAND IS ',F9.6)

WRITE(6,906) K

906 FORMAT('0 BELOW ARE THE PREDICTOR NUMBERS AND COEFFICIENTS FOR '

\*, 'THE ',I3, ' PREDICTAND')

WRITE(6,907)(J,IS(J),B(K,J),J=1,L)

907 FORMAT(4(2I4,E15.6))

6000 CONTINUE

C HAVE YOU DONE ALL PREDICTAND GROUPS.

IF(NP-IG) 1000,3000,3000

3000 STOP

END

SUBROUTINE SCRNP(IP,MM,ZZ,YZ)

DIMENSION ZZ(131,131),YZ(131,131)

READ(1,900) ((ZZ(I,J),J=1,IP),I=1,IP)

C TRANSPOSE OF YZ IS NEEDED

READ(2,900) ((YZ(J,I),J=1,MM),I=1,IP)

900 FORMAT(145F7.0)

WRITE(6,901)

901 FORMAT('1 BELOW IS THE ZZ MATRIX')

WRITE(6,902) ((I,J,ZZ(I,J),J=1,IP),I=1,IP)

902 FORMAT(5(2I4,F10.0)/)

WRITE(6,903)

903 FORMAT('1 BELOW IS THE YZ MATRIX')

WRITE(6,902) ((I,J,YZ(I,J),J=1,IP),I=1,MM)

RETURN

END